

# LOCAL AND GLOBAL CONVERGENCE OF AN INERTIAL VERSION OF FORWARD-BACKWARD SPLITTING \*

PATRICK R. JOHNSTONE<sup>†</sup> AND PIERRE MOULIN<sup>†</sup>

**Abstract.** A problem of great interest in optimization is to minimize a sum of two closed, proper, and convex functions where one is smooth and the other has a computationally inexpensive proximal operator. In this paper we analyze a family of Inertial Forward-Backward Splitting (I-FBS) algorithms for solving this problem. We first apply a global Lyapunov analysis to I-FBS and prove weak convergence of the iterates to a minimizer in a real Hilbert space. We then show that the algorithms achieve local linear convergence for “sparse optimization”, which is the important special case where the nonsmooth term is the  $\ell_1$ -norm. This result holds under either a restricted strong convexity or a strict complimentary condition and we do not require the objective to be strictly convex. For certain parameter choices we determine an upper bound on the number of iterations until the iterates are confined on a manifold containing the solution set and linear convergence holds.

The local linear convergence result for sparse optimization holds for the Fast Iterative Shrinkage and Soft Thresholding Algorithm (FISTA) due to Beck and Teboulle which is a particular parameter choice for I-FBS. In spite of its optimal global objective function convergence rate, we show that FISTA is not optimal for sparse optimization with respect to the local convergence rate. We determine the locally optimal parameter choice for the I-FBS family. Finally we propose a method which inherits the excellent global rate of FISTA but also has excellent local rate.

**Key words.** proximal gradient methods, forward-backward splitting, inertial methods,  $\ell_1$ -regularization, local linear convergence

**AMS subject classifications.** 65K05, 65K15, 90C06, 90C25

**1. Introduction.** We are concerned with the following important problem:

$$\underset{x \in \mathcal{H}}{\text{minimize}} F(x) = f(x) + g(x), \quad (1.1)$$

where  $\mathcal{H}$  is a Hilbert space over the real numbers, the functions  $f : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  and  $g : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  are proper, convex and closed, and in addition  $f$  is Gâteaux differentiable, and has a Lipschitz continuous gradient. Problems of this form have come under considerable attention in recent years in applications such as machine learning [1, 2], compressed sensing [3, 4] and image processing [5, 6] among many other examples. Of particular interest in this paper will be the special case which we will call sparse optimization (SO).

$$(\text{Problem SO}) \quad \underset{x \in \mathbb{R}^n}{\text{minimize}} F(x) = f(x) + \rho \|x\|_1,$$

where  $\rho \geq 0$ , and  $\|x\|_1 = \sum_{i=1}^n |x_i|$ . We refer to this problem as “sparse optimization” because the  $\ell_1$ -norm encourages sparse solutions. When  $f(x) = \frac{1}{2} \|b - Ax\|^2$  with  $A \in \mathbb{R}^{m \times n}$  and  $b \in \mathbb{R}^m$ , Problem SO is often referred to as sparse least squares (Problem  $\ell_1$ -LS), basis pursuit denoising or LASSO. This problem is of central importance in compressed sensing and also has applications in machine learning [7] and image processing [8]. Other important instances of Problem (1.1) include least squares with a total-variation [9] or nuclear-norm [10] regularizer, and minimization constrained to a closed and convex set.

\*The proofs of Thms. 4.1, 5.1, 5.2, and 5.6 of this manuscript contain several errors. These errors have been fixed in a revised and rewritten manuscript entitled “Local and Global Convergence of a General Inertial Proximal Splitting Scheme” arxiv id. 1602.02726. We recommend reading this updated manuscript.

<sup>†</sup>Beckman Institute, University of Illinois, 405 N. Mathews Ave., Urbana, IL, 61801, USA (contact: prjohns2@illinois.edu)

**1.1. Background.** In this paper we focus on *first-order splitting methods* for solving Problem (1.1). These methods use evaluations of  $F$ , gradients of the smooth part  $f$ , and evaluations of the *proximal operator* of the nonsmooth part  $g$ . In particular we focus on the *forward-backward splitting algorithm* (FBS), which is a classical first-order splitting approach to solving Problem (1.1) [11, 12]. In fact FBS was developed for the more general monotone inclusion problem which includes Problem (1.1) as a special case. FBS involves a “forward” step, which is an explicit gradient step with respect to the differentiable component  $f$  and a “backward” step, which is an implicit, proximal step with respect to  $g$ . For many popular instances of  $g$  this proximal step is computationally inexpensive [13]. The convergence rate of the objective function to the infimum is  $O(1/k)$ , which is better than the  $O(1/\sqrt{k})$  rate achieved by the “black-box” subgradient method, and is the same as if the possibly nonsmooth component were not present. Weak convergence of the iterates is also guaranteed and linear convergence occurs on strongly convex problems [14]. FBS is also commonly referred to as the proximal gradient method [15] and for the special case of Problem SO, it is known as the iterative shrinkage and soft thresholding algorithm (ISTA) owing to the form of the proximal step w.r.t. the  $\ell_1$ -norm [16, 17, 18]. Other first-order splitting methods include ADMM [19], linearized and preconditioned ADMM [20], primal-dual methods [21], Bregman iterations [22] and generalized FBS [23]. These methods can deal with more complicated situations such as when  $g$  is composed with a bounded linear operator or when the sum of  $m > 1$  proximable<sup>1</sup> functions is present.

Nesterov developed several methods for minimizing a convex function with Lipschitz gradient ([24], [25] chapter 2). These methods obtain the best objective function convergence rate possible by any first order method. Specifically, they guarantee a convergence rate of  $O(1/k^2)$  for the objective function, which is optimal in the worst case sense for convex functions with Lipschitz gradient. Note that this improves the  $O(1/k)$  rate achieved by classical gradient descent.

In [16], Beck and Teboulle extended Nesterov’s method of [25] to Problem (1.1), allowing for the presence of the possibly nonsmooth function  $g$ . Their method, FISTA, combines Nesterov’s inertial update into an FBS framework using the same sequence of “momentum” parameters. FISTA corresponds to a particular parameter choice for the following suite of algorithms, which we will call Inertial Forward-Backward Splitting (I-FBS).

$$(\text{I-FBS}) : \forall k \in \mathbb{N}, \quad \begin{cases} y^{k+1} = x^k + \alpha_k(x^k - x^{k-1}) \\ x^{k+1} = \text{prox}_{\lambda_k g}(y^{k+1} - \lambda_k \nabla f(y^{k+1})) \end{cases}$$

with  $x^0, x^1 \in \mathcal{H}$  chosen arbitrarily (typically  $x^0 = x^1$ ). The sequences  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$  are a subset of  $\mathbb{R}_{\geq 0}$ . The proximal operator  $\text{prox}_g : \mathcal{H} \rightarrow \mathcal{H}$  will be properly defined in Section 2.2. Beck and Teboulle showed that for a specific choice of  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$ , I-FBS obtains the optimal  $O(1/k^2)$  rate in terms of the objective function, however they did not prove convergence of the iterates  $\{x^k\}_{k \in \mathbb{N}}$  to a minimizer which is also unknown for Nesterov’s method. Tseng [26] showed that other choices also achieve  $O(1/k^2)$  rate. Recently in [27], Chambolle and Dossal considered a very similar choice of the parameters to Beck and Teboulle which obtains  $O(1/k^2)$  rate in the objective function and also weak convergence of the iterates to a minimizer. Throughout the rest of the paper we will refer to all these parameter choices for I-FBS

---

<sup>1</sup>Possessing a simple proximal operator.

that obtain the  $O(1/k^2)$  objective function rate as *FISTA-like* choices. Note that FBS corresponds to I-FBS with  $\alpha_k$  set to 0 for all  $k \in \mathbb{N}$  and  $\lambda_k$  in the range  $(0, 2/L)$  where  $L$  is the Lipschitz constant of  $\nabla f$ . Nesterov's method of [25] corresponds to I-FBS with the same parameter choice as FISTA and with  $g = 0$ .

One of the aims of this paper is to establish broad conditions for the convergence of the iterates of I-FBS to a minimizer of Problem (1.1). A generalization of the I-FBS family has been studied previously in [28] in the setting of monotone operator inclusion problems. However our global analysis proves convergence for a wider range of parameter choices than was proved there. An algorithm similar to I-FBS was developed in [29] for the more general problem of finding a fixed-point of a nonexpansive operator. However the conditions for convergence are far more strict than those developed in this paper. To the best of our knowledge the conditions for weak convergence of the iterates of I-FBS developed in this paper are novel in the literature. A more detailed comparison with existing literature is given in Section 3.

It has been observed that for the special case of Problem SO FBS exhibits *local linear convergence* (see e.g. [17, 18, 30, 31]), elsewhere called *eventual linear convergence* [32]. By this it is meant that there exists some  $N > 0$  such that for all  $k > N$  the iterates  $x^k$  are confined to a manifold containing the solution set and convergence to a solution is linear. It is not known whether I-FBS (including the FISTA-like choices) obtains local linear convergence for Problem SO, however recently [33] has made progress for the special case of Problem  $\ell_1$ -LS. In this paper, we address this by establishing local linear convergence of I-FBS for Problem SO for a broad range of parameter choices including the FISTA-like choices. Of course local linear convergence of the sequence  $\{x^k\}_{k \in \mathbb{N}}$  implies convergence of the entire sequence.

**1.2. Contributions of this Paper.** In the first part of the paper, we analyze I-FBS with an appropriate multi-step Lyapunov function. This approach allows us to develop novel conditions on the algorithmic parameters that imply convergence of the iterates to a minimizer (weak convergence in a real Hilbert space, ordinary convergence in  $\mathbb{R}^n$ ). This widens the range of possible parameter choices beyond those proposed in prior art such as [28].

In the second part of the paper, we consider in detail the behavior of I-FBS applied to Problem SO. We show that after a finite number of iterations I-FBS reduces to minimizing a local function on a reduced support subject to an orthant constraint. This result holds for the FISTA-like choices along with a wide range of other parameter choices. Next we show that a simple “locally optimal” parameter choice for I-FBS obtains a local linear convergence rate with the best asymptotic iteration complexity. The asymptotically optimal iteration complexity is better than that obtained by the FISTA-like choices and by ISTA. The improvement gained by I-FBS over ISTA when the correct amount of momentum is added is equivalent to the improvement that Nesterov's accelerated method [25] achieves over gradient descent for strongly convex functions with Lipschitz gradients. As a corollary of our analysis, we show that the adaptive momentum restart scheme proposed in [34] achieves the optimal iteration complexity. In contrast the analysis in [34] is only valid for strongly convex quadratic functions. Finally for parameter choices for which the “momentum parameter”  $\alpha_k$  is bounded away from 1, we determine an explicit upper bound on the number of iterations until convergence to the optimal manifold.

With little effort our analysis of I-FBS for Problem SO can be adapted to apply to the splitting inertial proximal method (SIPM) proposed by Moudafi and Oliny [35]. This method is a direct generalization of the heavy ball with friction method (HBF)

[36] to proximal splitting problems and differs from I-FBS in that the gradient of  $f$  is computed at  $x^k$  rather than  $y^{k+1}$ . We show that SIPM also achieves local linear convergence for this problem under appropriate parameter constraints.

The paper is organized as follows. In Section 2, notation and assumptions are discussed. In Section 3, we precisely define the I-FBS family and discuss known convergence results in more detail. In Section 4 we apply our Lyapunov analysis to I-FBS. In Section 5 we derive convergence results for Problem SO. Finally, numerical experiments are presented in Section 6.

## 2. Preliminaries.

**2.1. Notation and Definitions.** Throughout the paper,  $\mathcal{H}$  is a Hilbert space over the field of real numbers,  $\langle \cdot, \cdot \rangle$  is the inner product and  $\| \cdot \|$  is the associated norm. Let  $\Gamma_0(\mathcal{H})$  be the set of all closed, convex and proper functions whose domain is a subset of  $\mathcal{H}$  and range is a subset of  $\mathbb{R} \cup \{+\infty\}$ . For any  $g : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  and point  $x \in \mathcal{H}$ , we denote by  $\partial_\epsilon g(x)$  for  $\epsilon \geq 0$  the  $\epsilon$ -enlargement of the *subdifferential*, defined as the set

$$\partial_\epsilon g(x) \triangleq \{v \in \mathcal{H} : g(y) \geq g(x) + \langle v, y - x \rangle - \epsilon, \forall y \in \mathcal{H}\} \quad (2.1)$$

which is always convex and closed and may be empty. We will use  $\partial g$  to denote  $\partial_0 g$ . When  $\partial g(x)$  is a singleton we will call it the *gradient* at  $x$ , denoted by  $\nabla g(x)$ .

For  $a : \mathbb{R} \rightarrow \mathbb{R}$  and  $b : \mathbb{R} \rightarrow \mathbb{R}$ , the notation  $a(k) = O(b(k))$  (resp.  $a(k) = \Omega(b(k))$ ) means there exists a constant  $C \geq 0$  such that  $\lim_{k \rightarrow \infty} a(k)/b(k) \leq C$  (resp.  $\lim_{k \rightarrow \infty} a(k)/b(k) \geq C$ ). The notation  $a(k) = o(b(k))$  means  $\lim_{k \rightarrow \infty} a(k)/b(k) = 0$ . We will say a sequence  $\{x^k\}_{k \in \mathbb{N}} \subset \mathcal{H}$  converges linearly to  $x^* \in \mathcal{H}$  with rate of convergence  $q \in (0, 1)$ , if  $\|x^k - x^*\| = O(q^k)$ . To be precise we will occasionally refer to this as *asymptotic* or *local* linear convergence. Note that this is different from nonasymptotic, or global linear convergence with rate  $q$ , in which case there exists a  $C \geq 0$  such that  $\|x^k - x^*\| \leq Cq^k$  for all  $k \in \mathbb{N}$ . In contrast local linear convergence allows for a finite number of iterations where such a relationship does not hold.

Define the *optimal value* of Problem (1.1) as

$$F^* \triangleq \inf_{x \in \mathcal{H}} F(x)$$

and the *solution set* as

$$X^* \triangleq \{x \in \mathcal{H} : F(x) = F^*\}.$$

Given a function  $a : \mathbb{R} \rightarrow \mathbb{R}$ , we say that the *iteration complexity* of a method for minimizing  $F$  is  $\Omega(a(\epsilon))$  if  $k = \Omega(a(\epsilon))$  implies  $F(x^k) - F^* = O(\epsilon)$ . To be precise we will occasionally refer to this as the asymptotic iteration complexity.

For a matrix  $A \in \mathbb{R}^{m \times n}$  and a set  $S \subset \{1, 2, \dots, n\}$ ,  $A_S$  will denote the matrix in  $\mathbb{R}^{m \times |S|}$  formed by taking the columns corresponding to the elements of  $S$ . For a vector  $v \in \mathbb{R}^n$ ,  $v_S$  will denote the  $|S| \times 1$  vector with entries given by the entries of  $v$  on the indices corresponding to the elements of  $S$ , and  $(v_S, 0)$  will denote the vector in  $\mathbb{R}^n$  equal to  $v$  on the indices corresponding to  $S$  and equal to zero everywhere else. Given  $c \in \mathbb{R}$  and  $x \in \mathbb{R}^n$ ,  $\text{sgn}(c)$  is defined as  $+1$  if  $c \geq 0$  and  $-1$  if  $c < 0$ ,  $\text{sgn}(x)$  is simply applying  $\text{sgn}(\cdot)$  element-wise. We will use the notation  $[c]_+ = \max(c, 0)$ .

**2.2. Proximal Operators.** The proximal operator  $\text{prox}_g : \mathcal{H} \rightarrow \mathcal{H}$  w.r.t. a function  $g \in \Gamma_0(\mathcal{H})$  is defined implicitly by

$$y - \text{prox}_g(y) \in \partial g(\text{prox}_g(y)),$$

and explicitly by

$$\text{prox}_g(y) = \arg \min_x \left\{ \frac{1}{2} \|x - y\|^2 + g(x) \right\}. \quad (2.2)$$

Since the function being minimized in (2.2) is strongly convex and in  $\Gamma_0(\mathcal{H})$ ,  $\text{prox}_g(y)$  exists and is unique for every  $y \in \mathcal{H}$  thus it is a well defined mapping with domain equal to  $\mathcal{H}$ . To be more general we will actually use the  $\epsilon$ -enlarged proximal operator, which is the set

$$\text{prox}_g^\epsilon(y) = \{v : y - v \in \partial_\epsilon g(v)\},$$

which is not necessarily uniquely defined (except when  $\epsilon = 0$ ). Note that  $\text{prox}_g(y) \in \text{prox}_g^\epsilon(y)$  for all  $\epsilon \geq 0$ . The use of  $\text{prox}_g^\epsilon$  allows for some approximation error in the computation of the proximal operator.

**2.3. Cocoercivity and Convexity.** We say that a Gâteaux differentiable and convex function  $f$  has a  $\frac{1}{L}$ -cocoercive gradient with  $L > 0$ , if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2, \quad \forall x, y \in \mathcal{H}. \quad (2.3)$$

Note this is equivalent to the gradient being  $L$ -Lipschitz continuous, i.e.

$$\|\nabla f(y) - \nabla f(x)\| \leq L \|y - x\|, \quad \forall x, y \in \mathcal{H}, \quad (2.4)$$

For a proof see [37] Lemma 1.4 and the Baillon-Haddad Theorem [38]. We will need the following two standard properties of such a function. For all  $u, v \in \mathcal{H}$ :

$$f(u) - f(v) \leq \langle \nabla f(v), u - v \rangle + \frac{L}{2} \|u - v\|^2, \quad (2.5)$$

and (by convexity)

$$f(u) - f(v) \leq \langle \nabla f(u), u - v \rangle. \quad (2.6)$$

We are now ready to formally state our Assumptions for Problem (1.1).

**Assumption 1.**  $f$  and  $g$  are in  $\Gamma_0(\mathcal{H})$ ,  $f$  is Gâteaux differentiable everywhere and has a  $1/L$ -cocoercive gradient with  $L > 0$ , and  $F^* > -\infty$ .

**2.4. Properties of Sparse Optimization.** We now outline our assumptions for Problem SO and discuss some of its properties.

**Assumption SO.**  $f \in \Gamma_0(\mathcal{H})$ , is twice differentiable everywhere, and has a  $1/L$ -cocoercive gradient with  $L > 0$ .  $F^* > -\infty$  and  $X^*$  is non-empty.

The main difference between Assumption SO and Assumption 1 is that we additionally assume that  $f$  is twice differentiable. Let  $H(x)$  denote the Hessian of  $f$  at  $x$ . Then the Lipschitz constant  $L$  of the gradient is equal to the supremum of the largest

eigenvalue of  $H(x)$  over all  $x$ . Furthermore note that  $\|\cdot\|_1$  is in  $\Gamma_0(\mathcal{H})$ . Finally note that for  $\rho > 0$  the function  $f(x) + \rho\|x\|_1$  is coercive thus  $X^*$  is non-empty.

Problem SO includes Problem  $\ell_1$ -LS, defined as

$$(\text{Problem } \ell_1\text{-LS}) \quad \underset{x \in \mathbb{R}^n}{\text{minimize}} \quad F(x) = \frac{1}{2}\|b - Ax\|^2 + \rho\|x\|_1,$$

where  $A \in \mathbb{R}^{m \times n}$  and  $b \in \mathbb{R}^m$ . The solution set  $X^*$  of Problem  $\ell_1$ -LS is always non-empty. The function  $f$  has gradient equal to  $A^T(Ax - b)$  which is Lipschitz-continuous with Lipschitz constant  $L$  equal to the largest eigenvalue of  $A^T A$ .

The proximal operator associated with  $\rho\|\cdot\|_1$  is the *shrinkage and soft-thresholding* operator  $S_\rho(v) : \mathbb{R} \rightarrow \mathbb{R}$ , applied element-wise. It is defined as  $S_\rho(v) \triangleq [|v| - \rho]_+ \text{sgn}(v)$ , and thus

$$\{\text{prox}_{\rho\|\cdot\|_1}(z)\}_i = S_\rho(z_i), \quad i = 1, 2, \dots, n. \quad (2.7)$$

In the analysis of I-FBS applied to Problem SO we will need the following result proved in [17].

**THEOREM 2.1** (Theorem 2.1 [17]). *For problem SO suppose Assumption SO holds, then there exists a vector  $h^* \in \mathbb{R}^n$  such that for all  $x^* \in X^*$ ,  $\nabla f(x^*) = h^*$ . Furthermore, for all  $i \in \{1, 2, \dots, n\}$ ,*

$$\frac{h_i^*}{\rho} \begin{cases} = -1 & \text{if } \exists x \in X^* : x_i > 0 \\ = +1 & \text{if } \exists x \in X^* : x_i < 0 \\ \in [-1, 1] & \text{else.} \end{cases}$$

The following two sets also used in [17] will also be crucial to our analysis. Let  $D \triangleq \{i : |h_i^*| < \rho\}$  and  $E \triangleq \{i : |h_i^*| = \rho\}$ . Note that  $D \cap E = \emptyset$  and  $D \cup E = \{1, 2, \dots, n\}$ . By Theorem 2.1, we can infer that  $\text{supp}(x^*) \subseteq E$  for all  $x^* \in X^*$ . Finally, define

$$\omega \triangleq \min\{\rho - |h_i^*| : i \in D\} > 0.$$

We will need the following Lemma proved in [17].

**LEMMA 2.2** (Lemma 4.1 [17]). *Under Assumption SO, if  $\lambda \in [0, 2/L)$ ,*

$$\|x - \lambda \nabla f(x) - (y - \lambda \nabla f(y))\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

An alternative definition of cocoercivity is to say that if an operator  $T : \mathcal{H} \rightarrow \mathcal{H}$  is  $\gamma$ -cocoercive than  $\gamma T$  is *firmly nonexpansive*. Thus Lemma 2.2 is just an elementary property of firmly nonexpansive operators (see Proposition 4.2 (iii), and Proposition 4.33 [39]).

Finally, the following properties of  $S_\nu$  will be useful.

**LEMMA 2.3** (Lemma 3.2 [17]). *Fix any  $a$  and  $b$  in  $\mathbb{R}$ , and  $\nu \geq 0$ :*

- *The function  $S_\nu$  is nonexpansive. That is,*

$$|S_\nu(a) - S_\nu(b)| \leq |a - b|.$$

- *If  $|b| \geq \nu$  and  $\text{sgn}(a) \neq \text{sgn}(b)$  then*

$$|S_\nu(a) - S_\nu(b)| \leq |a - b| - \nu.$$

- *If  $S_\nu(a) \neq 0 = S_\nu(b)$  then  $|a| > \nu, |b| < \nu$  and*

$$|S_\nu(a) - S_\nu(b)| \leq |a - b| - (\nu - |b|). \quad (2.8)$$

**3. I-FBS.** To be more general, our global analysis will apply to the following I-FBS family.

$$(I\text{-FBS-}\epsilon) : \forall k \in \mathbb{N} \quad \begin{cases} y^{k+1} = x^k + \alpha_k(x^k - x^{k-1}) \\ x^{k+1} \in \text{prox}_{\lambda_k g}^{\epsilon_k}(y^{k+1} - \lambda_k \nabla f(y^{k+1})) \end{cases}$$

with  $x^0, x^1 \in \mathcal{H}$  chosen arbitrarily. Note that for any  $\epsilon_k \geq 0$ ,

$$\text{prox}_{\lambda_k g}(y^{k+1} - \lambda_k \nabla f(y^{k+1})) \in \text{prox}_{\lambda_k g}^{\epsilon_k}(y^{k+1} - \lambda_k \nabla f(y^{k+1})).$$

We will refer to  $\{\alpha_k\}_{k \in \mathbb{N}}$  as the “momentum” parameters and  $\{\lambda_k\}_{k \in \mathbb{N}}$  as the “step-size” parameters. The algorithm differs from I-FBS in that it uses the  $\epsilon$ -enlarged sub-differential, allowing for some error in the computation of the proximal operator.

**3.1. Known Convergence Results.** Beck and Teboulle [16] proposed the following choice of parameters for I-FBS (I-FBS- $\epsilon$  with the  $\epsilon_k$  set to 0 for all  $k \in \mathbb{N}$ ),

$$\forall k \in \mathbb{N}, \lambda_k = \frac{1}{L}, \alpha_k = \frac{t_k - 1}{t_{k+1}}, \text{ where } t_{k+1} = \frac{1 + \sqrt{4t_k^2 + 1}}{2}, t_1 = 1. \quad (3.1)$$

The method is known as FISTA. With this choice of parameters, Beck and Teboulle showed that the objective function converges to the minimum at the worst-case optimal rate of  $O(1/k^2)$ . In fact the  $O(1/k^2)$  rate holds for a variety of choices of  $\{\alpha_k\}_{k \in \mathbb{N}}$  which all have the form:  $\alpha_k = 1 - O(1/k)$  [26]. However the choice in (3.1) guarantees the largest possible decrease in a given upper bound of  $F(x^k)$  at each iteration. Chambolle and Dossal [27] considered I-FBS with a similar choice of  $\{\alpha_k\}_{k \in \mathbb{N}}$  to what was proposed by Beck and Teboulle. They investigated, for some  $a > 2$ ,

$$\forall k \in \mathbb{N}, 0 < \lambda_k \leq \frac{1}{L}, \alpha_k = \frac{t_k - 1}{t_{k+1}}, \text{ where } t_{k+1} = \frac{k + a - 1}{a}, t_1 = 1. \quad (3.2)$$

With this choice of parameters, the authors showed that the objective function achieves the optimal  $O(1/k^2)$  convergence rate and in addition  $\{x^k\}_{k \in \mathbb{N}}$  weakly converges to a minimizer.

In contrast to [16] and [27], our analysis establishes weak convergence of the iterates for a wide range of parameter choices. Indeed, the momentum sequence is not constrained to follow a particular sequence relationship, but instead must be constrained to  $\alpha_k \in [0, 1]$  and  $\limsup \alpha_k < 1$ . However we do not guarantee the  $O(1/k^2)$  objective function rate.

Lorenz and Pock [28] generalized I-FBS to the problem of finding a zero of the sum of two maximal monotone operators  $A$  and  $B$ , one of which is cocoercive. Setting  $A = \nabla f$  and  $B = \partial g$  recovers Problem (1.1). They also replaced the scalar step-size  $\lambda_k$  with a general positive definite operator  $\lambda_k M^{-1}$ . Lorenz and Pock proved weak convergence of the iterates to a solution provided certain restrictions on  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$ . The restrictions on  $\alpha_k$  are stronger than those derived in our global analysis. In their analysis, if the step-size  $\lambda_k$  is fixed to  $1/L$ ,  $\alpha_k$  is restricted to be less than  $\sqrt{5} - 2 \approx 0.24$ , whereas, as we shall see in Section 4, our Lyapunov analysis allows  $\alpha_k \in [0, 1]$ , so long as  $\limsup \alpha_k < 1$ . For the step-size, their conditions are less restrictive than ours, allowing for values of  $\lambda_k$  up to  $2/L$ , whereas our analysis only allows up to  $1/L$ . However in their analysis larger values of  $\lambda_k$  lead to a smaller range of feasible values for  $\alpha_k$  reducing to 0 as  $\lambda_k$  approaches  $2/L$ .

In [29], an inertial version of the classical Krasnosel'skii-Mann (KM) algorithm was analyzed. The KM algorithm finds the fixed points of a nonexpansive operator

$T : \mathcal{H} \rightarrow \mathcal{H}$ . Setting the operator  $T(x) = \text{prox}_{\lambda_k g}(x - \lambda_k \nabla f(x))$  in the inertial KM method of [29] recovers I-FBS, since a point  $x$  is a fixed point of  $T$  if and only if it is a solution of Problem (1.1). The analysis of [29] proves weak convergence of the iterates to a fixed point but relies on verifying conditions of the form:  $\sum \alpha_k \|x^k - x^{k-1}\|^p < \infty$  with  $p$  equal to 1 and 2. In general this condition must be enforced online, restricting the range of possible choices for the sequence of momentum parameters. However, it was shown in [40] that choosing  $\alpha_k$  to be nondecreasing and satisfying  $\alpha_k \in [0, \bar{\alpha})$  with  $\bar{\alpha} < 1/3$  suffices to ensure the condition is satisfied and thus prove weak convergence. This condition is more restrictive than the ones derived in this paper for the special case of Problem (1.1).

**3.2. Known Convergence Results for Sparse Optimization.** The FISTA-like sequences for  $\{\alpha_k\}_{k \in \mathbb{N}}$  defined in (3.1) and (3.2) both converge to 1. As we will see in Section 5.5 this is not desirable for Problem SO. In the language of dynamical systems, when the momentum is too high the iterates move into an “underdamped regime” leading to oscillations in the objective function and slow convergence (see [34] for an analysis in the strongly-convex quadratic case). We will show that for Problem SO the FISTA-like choices are not optimal from the viewpoint of asymptotic rate of convergence under a local strong-convexity assumption (Corollary 5.4) or a strict complementarity condition (Corollary 5.5).

In [33], the behavior of ISTA and FISTA (i.e. I-FBS with parameter choice (3.1)) applied to Problem  $\ell_1$ -LS was investigated through a spectral analysis. The authors show that both algorithms obtain local linear convergence for this problem, under the condition that the minimizer is unique, but without an estimate for the number of iterations until convergence to the optimal manifold. Furthermore they determine that the local rate of convergence of FISTA is worse than ISTA, while the transient behavior of FISTA is better than ISTA. Therefore they suggest switching from FISTA to ISTA once the optimal manifold has been identified. Our contribution differs in several ways. We note that the poor local performance of the FISTA-like choices is due to having the momentum parameter converge to 1. Therefore we determine the optimal value for the momentum parameter that should be used in the asymptotic regime which allows for a better asymptotic rate than both ISTA and FISTA and suggest a heuristic method for estimating the optimal momentum. We also show that the adaptive restart method of [34] will achieve the  $O(1/k^2)$  rate in the transient regime and the optimal asymptotic rate. Furthermore our analysis holds for Problem SO with Problem  $\ell_1$ -LS as a special case and we do not require the minimizer to be unique. Finally, in the case where  $\limsup \alpha_k < 1$ , we provide explicit upper bounds on the number of iterations until I-FBS has converged to the optimal manifold.

In [41] a method was developed for solving Problem (1.1) when  $f$  is strongly convex. The method is equivalent to I-FBS with the same prescription for  $\{\alpha_k\}_{k \in \mathbb{N}}$  as determined by Nesterov for his method for minimizing strongly convex functions (constant scheme 2.2.8. of [25]). However it also includes a backtracking procedure for adjusting  $\{\lambda_k\}_{k \in \mathbb{N}}$  and  $\{\alpha_k\}_{k \in \mathbb{N}}$  when the strong convexity and Lipschitz gradient parameters are not known. The authors of [41] also extended their method to Problem  $\ell_1$ -LS including the case where  $f$  is not strongly convex. The authors showed that under conditions on the matrix  $A$  related to the Restricted Isometry Property (RIP) used in compressed sensing, their algorithm obtains nonasymptotic (global) linear convergence, so long as the initial vector is sufficiently sparse. However, as the authors note the RIP-like conditions are much stronger than those typically found in the literature. Indeed the conditions are much stronger than those required in



our proof of local linear convergence. We establish that I-FBS obtains local linear convergence regardless of the initialization point. Furthermore no RIP-like assumptions are necessary. Local linear convergence can be proved under the mild condition that the smallest eigenvalue of the Hessian restricted to the support of a minimum is non-zero at the minimum point. Or if this does not hold, under a common strict-complementarity condition (see Section 5.4). That being said, it should be noted that local linear convergence is not as strong a statement as global linear convergence

**4. A Global Analysis of I-FBS.** This section derives conditions on  $\{\epsilon_k\}_{k \in \mathbb{N}}$ ,  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$  which imply weak convergence of the iterates  $\{x^k\}_{k \in \mathbb{N}}$  of I-FBS to a minimizer of Problem (1.1). Throughout the rest of the paper, let  $\Delta_{k+1}$  denote  $x^{k+1} - x^k$ . Given  $S, T \subset \mathcal{H}$ , define  $d(S, T) \triangleq \min_{s \in S, t \in T} \|s - t\|$ .

**THEOREM 4.1.** *Suppose that Assumption 1 holds. Assume  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and satisfies  $0 < \lambda_k \leq 1/L$  for all  $k$ , and  $\{\epsilon_k\}_{k \in \mathbb{N}}$  satisfies  $\epsilon_k \geq 0$  for all  $k$  and  $\sum_k \epsilon_k < \infty$ . If  $0 \leq \alpha_k \leq 1$  for all  $k$  and  $\limsup \alpha_k < 1$ , then for the iterates of I-FBS- $\epsilon$ , we have*

- (i)  $\sum_{k=1}^{\infty} \|\Delta_k\|^2 < \infty$ .
- (ii)  $\lim_{k \rightarrow \infty} d(0, \nabla f(x^k) + \partial_{\epsilon_k} g(x^k)) = \lim_{k \rightarrow \infty} d(0, \nabla f(y^k) + \partial_{\epsilon_k} g(y^k)) = 0$ .
- (iii) If, in addition,  $X^*$  is non-empty, then  $x^k$  converges weakly to some  $\hat{x} \in X^*$ .

*Proof.* The proof consists of two parts. In the first, we prove statements (i) and (ii) using arguments inspired by Alvarez' analysis of the inertial proximal method in [42]. In the second part, we invoke Opial's Lemma [43] to prove statement (iii). The second part is inspired by the analysis of the splitting inertial proximal algorithm by Moudafi and Oliny in [35].

**Proof of statements (i) and (ii).** Define the *Lyapunov function*, or discrete energy, to be

$$E_k \triangleq \frac{\alpha_k}{2\lambda_k} \|\Delta_k\|^2 + f(x^k) + g(x^k).$$

Note that this is the same energy function used by Alvarez [44]. Inequalities (2.5), (2.6) and (2.1) imply

$$\begin{aligned} E_{k+1} - E_k &= \frac{\alpha_{k+1}}{2\lambda_{k+1}} \|\Delta_{k+1}\|^2 - \frac{\alpha_k}{2\lambda_k} \|\Delta_k\|^2 + f(x^{k+1}) - f(x^k) + g(x^{k+1}) - g(x^k) \\ &\leq \frac{\alpha_{k+1}}{2\lambda_{k+1}} \|\Delta_{k+1}\|^2 - \frac{\alpha_k}{2\lambda_k} \|\Delta_k\|^2 + \langle \nabla f(y^{k+1}) + v, \Delta_{k+1} \rangle \\ &\quad + \frac{L}{2} \|x^{k+1} - y^{k+1}\|^2 + \epsilon_k, \quad \forall v \in \partial_{\epsilon_k} g(x^{k+1}). \end{aligned} \tag{4.1}$$

Note that the existence of a subgradient  $v$  is guaranteed because the  $\epsilon$ -enlarged proximal operator has domain equal to  $\mathcal{H}$ . Using (4.1), the  $x^{k+1}$  - update in I-FBS and

the fact that  $\lambda_k \leq \lambda_{k+1}$ , we infer that

$$\begin{aligned}
E_{k+1} - E_k &\leq \frac{\alpha_{k+1}}{2\lambda_{k+1}} \|\Delta_{k+1}\|^2 - \frac{\alpha_k}{2\lambda_k} \|\Delta_k\|^2 - \frac{1}{\lambda_k} \langle x^{k+1} - y^{k+1}, \Delta_{k+1} \rangle \\
&\quad + \frac{L}{2} \|\Delta_{k+1} - \alpha_k \Delta_k\|^2 + \epsilon_k \\
&= \frac{\alpha_{k+1}}{2\lambda_{k+1}} \|\Delta_{k+1}\|^2 - \frac{\alpha_k}{2\lambda_k} \|\Delta_k\|^2 - \frac{1}{\lambda_k} \langle \Delta_{k+1} - \alpha_k \Delta_k, \Delta_{k+1} \rangle \\
&\quad + \frac{L}{2} (\|\Delta_{k+1}\|^2 + \alpha_k^2 \|\Delta_k\|^2) - \alpha_k L \langle \Delta_{k+1}, \Delta_k \rangle + \epsilon_k \\
&\leq \left( \frac{L}{2} - \frac{1}{\lambda_k} + \frac{\alpha_{k+1}}{2\lambda_k} \right) \|\Delta_{k+1}\|^2 + \left( \frac{\alpha_k^2 L}{2} - \frac{\alpha_k}{2\lambda_k} \right) \|\Delta_k\|^2 \\
&\quad + \frac{\alpha_k(1 - \lambda_k L)}{\lambda_k} \langle \Delta_{k+1}, \Delta_k \rangle + \epsilon_k \\
&= -\frac{\alpha_k(1 - \lambda_k L)}{2\lambda_k} \|\Delta_{k+1} - \Delta_k\|^2 - \frac{2 + \lambda_k L(\alpha_k - 1) - \alpha_k - \alpha_{k+1}}{2\lambda_k} \|\Delta_{k+1}\|^2 \\
&\quad - \frac{L\alpha_k(1 - \alpha_k)}{2\lambda_k} \|\Delta_k\|^2 + \epsilon_k.
\end{aligned}$$

Moving terms to the other side and summing implies, for all  $N \in \mathbb{Z}_+$ ,

$$\begin{aligned}
&\sum_{k=1}^N \left[ \frac{\alpha_k(1 - \lambda_k L)}{2\lambda_k} \|\Delta_{k+1} - \Delta_k\|^2 + \frac{2 + \lambda_k L(\alpha_k - 1) - \alpha_k - \alpha_{k+1}}{2\lambda_k} \|\Delta_{k+1}\|^2 \right. \\
&\quad \left. + \frac{L\alpha_k(1 - \alpha_k)}{2\lambda_k} \|\Delta_k\|^2 \right] \\
&\leq E_1 - E_{N+1} + \sum_{k=1}^N \epsilon_k \\
&\leq E_1 - F^* + \sum_{k=1}^N \epsilon_k = F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 + \sum_{k=1}^N \epsilon_k < \infty. \tag{4.2}
\end{aligned}$$

Inequality (4.2) along with the assumptions on  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$  imply statement (i). Statement (i) implies  $\|\Delta_{k+1}\| \rightarrow 0$ , therefore  $\|x^k - y^{k+1}\| \rightarrow 0$  via the  $y^{k+1}$ -update of I-FBS. This implies that  $\|x^k - y^k\| \rightarrow 0$ , because  $\|x^k - y^k\| \leq \|x^{k-1} - y^k\| + \|x^{k-1} - x^k\|$ . Finally, using the  $x^{k+1}$ -update of I-FBS we infer that

$$\lim_{k \rightarrow \infty} d(\nabla f(y^k) + \partial_{\epsilon_k} g(y^k), 0) = \lim_{k \rightarrow \infty} d(\nabla f(x^k) + \partial_{\epsilon_k} g(x^k), 0) = 0. \tag{4.3}$$

**Proof of statement (iii).** If  $x^{v_k}$  is a subsequence which weakly converges to  $x'$ , then the  $y^{k+1}$ -update of I-FBS implies  $y^{v_k}$  also weakly converges to  $x'$ . This, combined with the  $x^{k+1}$ -update implies that  $x' \in X^*$ . Suppose that for any  $x^* \in X^*$ , the sequence  $\|x^k - x^*\|$  has a limit. This implies the sequence  $x^k$  is bounded and therefore it has at least one weakly-convergent subsequence,  $x^{\nu_k}$  (ordinary convergence in  $\mathbb{R}^n$ ). By the above reasoning the limit of this subsequence,  $\tilde{x}$  must be in  $X^*$ . Furthermore  $\lim_k \|x^k - \tilde{x}\|$  exists. Consider another subsequence  $x^{\nu'_k}$  which converges to  $\tilde{x}' \in X^*$ . By considering the fact that  $\lim_k \|x^{\nu_k} - \tilde{x}\|^2 = \lim_k \|x^{\nu'_k} - \tilde{x}\|^2$  and the corresponding statement for  $\tilde{x}'$ , one can see that  $\|\tilde{x} - \tilde{x}'\| = 0$ . Therefore the set of weakly convergent subsequences is the singleton  $\{\tilde{x}\}$ . Thus  $x^k$  weakly converges to  $\tilde{x} \in X^*$  (This is Opial's Lemma [43]).

Assume  $X^*$  is non-empty. We now proceed to show that, for any  $x^* \in X^*$ , the sequence  $\|x^k - x^*\|$  has a limit. Our proof closely follows Moudafi and Oliny's analysis [35], and is similar to the later variants [27, 28]. The main difference is we allow for  $\alpha_k$  to be 1 for a finite number of iterations. Fix  $x^* \in X^*$  and define  $\varphi_k = \frac{1}{2}\|x^k - x^*\|^2$ . Now

$$\varphi_k - \varphi_{k+1} = \frac{1}{2}\|\Delta_{k+1}\|^2 + \langle x^{k+1} - y^{k+1}, x^* - x^{k+1} \rangle + \alpha_k \langle \Delta_k, x^* - x^{k+1} \rangle. \quad (4.4)$$

Since

$$\begin{aligned} -x^{k+1} + y^{k+1} - \lambda_k \nabla f(y^{k+1}) &\in \lambda_k \partial_{\epsilon_k} g(x^{k+1}), \\ -\lambda_k \nabla f(x^*) &\in \lambda_k \partial_{\epsilon_k} g(x^*) \text{ and } \langle \partial_{\epsilon_k} g(x^{k+1}) - \partial_{\epsilon_k} g(x^*), x^{k+1} - x^* \rangle \geq -\epsilon, \text{ it follows that} \\ \langle x^{k+1} - y^{k+1} + \lambda_k (\nabla f(y^{k+1}) - \nabla f(x^*)), x^* - x^{k+1} \rangle &\geq -\lambda_k \epsilon_k. \end{aligned} \quad (4.5)$$

Combining (4.4) and (4.5) we obtain

$$\begin{aligned} \varphi_k - \varphi_{k+1} &\geq \frac{1}{2}\|\Delta_{k+1}\|^2 + \lambda_k \langle \nabla f(y^{k+1}) - \nabla f(x^*), x^{k+1} - x^* \rangle - \alpha_k \langle \Delta_k, x^{k+1} - x^* \rangle \\ &\quad - \lambda_k \epsilon_k. \end{aligned} \quad (4.6)$$

Now

$$\begin{aligned} \langle \Delta_k, x^{k+1} - x^* \rangle &= \langle \Delta_k, x^k - x^* \rangle + \langle \Delta_k, \Delta_{k+1} \rangle \\ &= \varphi_k - \varphi_{k-1} + \frac{1}{2}\|\Delta_k\|^2 + \langle \Delta_k, \Delta_{k+1} \rangle. \end{aligned} \quad (4.7)$$

Combining (4.6) and (4.7) yields

$$\begin{aligned} \varphi_{k+1} - \varphi_k - \alpha_k(\varphi_k - \varphi_{k-1}) &\leq -\frac{1}{2}\|\Delta_{k+1}\|^2 + \alpha_k \langle \Delta_k, \Delta_{k+1} \rangle + \frac{\alpha_k}{2}\|\Delta_k\|^2 \\ &\quad - \lambda_k \langle \nabla f(y^{k+1}) - \nabla f(x^*), x^{k+1} - x^* \rangle \\ &\quad + \lambda_k \epsilon_k. \end{aligned} \quad (4.8)$$

Now we use the fact that  $\nabla f$  is cocoercive as follows. Inequality (2.3) implies

$$\begin{aligned} \lambda_k \langle \nabla f(y^{k+1}) - \nabla f(x^*), x^{k+1} - x^* \rangle &= \lambda_k (\langle \nabla f(y^{k+1}) - \nabla f(x^*), y^{k+1} - x^* \rangle \\ &\quad + \langle \nabla f(y^{k+1}) - \nabla f(x^*), x^{k+1} - y^{k+1} \rangle) \\ &\geq \frac{\lambda_k}{L} (\|\nabla f(y^{k+1}) - \nabla f(x^*)\|^2 \\ &\quad + \langle \nabla f(y^{k+1}) - \nabla f(x^*), x^{k+1} - y^{k+1} \rangle) \\ &\geq -\frac{\lambda_k L}{4} \|x^{k+1} - y^{k+1}\|^2 \end{aligned} \quad (4.9)$$

where (4.9) follows by completing the square. Combining (4.8) and (4.9) we infer

$$\begin{aligned} \varphi_{k+1} - \varphi_k - \alpha_k(\varphi_k - \varphi_{k-1}) &\leq -\frac{1}{2}\|\Delta_{k+1}\|^2 + \alpha_k \langle \Delta_k, \Delta_{k+1} \rangle + \frac{\alpha_k}{2}\|\Delta_k\|^2 \\ &\quad + \frac{\lambda_k L}{4} \|\Delta_{k+1} - \alpha_k \Delta_k\|^2 + \lambda_k \epsilon_k \\ &= \frac{\alpha_k}{2} \left( \frac{\lambda L}{2} - 1 \right) \|\Delta_{k+1} - \Delta_k\|^2 \\ &\quad + \frac{\alpha_k}{4} (4 + \lambda L (\alpha_k - 1)) \|\Delta_k\|^2 \\ &\quad + \frac{\lambda L - 2}{4} (1 - \alpha_k) \|\Delta_{k+1}\|^2 + \lambda_k \epsilon_k. \end{aligned} \quad (4.10)$$

Note that the coefficients of  $\|\Delta_{k+1} - \Delta_k\|^2$  and  $\|\Delta_{k+1}\|^2$  are non-positive. Set  $\theta_k \triangleq \varphi_k - \varphi_{k-1}$  and

$$\delta_k \triangleq \frac{\alpha_k}{4} (4 + \lambda L(\alpha_k - 1)) \|\Delta_k\|^2 + \lambda_k \epsilon_k \quad (4.11)$$

and note that  $\sum_{k=1}^{\infty} \delta_k < \infty$ .

The argument from now on is basically identical to [35] except we allow for sequences  $\alpha_k$  which are equal to 1 for a finite number of  $k$ . Restate (4.10) as

$$\begin{aligned} \theta_{k+1} &\leq \alpha_k \theta_k + \delta_k \\ &\leq \alpha_k [\theta_k]_+ + \delta_k. \end{aligned} \quad (4.12)$$

Since  $\limsup \alpha_k < 1$ , there exists an integer  $K \geq 0$  and  $\bar{\alpha} \in [0, 1)$  such that  $\alpha_k \leq \bar{\alpha} < 1$  for all  $k > K$ . This and (4.12) imply that, for  $k > K$

$$[\theta_{k+1}]_+ \leq \bar{\alpha} [\theta_k]_+ + \delta_k.$$

Thus for  $k > K$

$$[\theta_{k+1}]_+ \leq \bar{\alpha}^{k-K} [\theta_K]_+ + \sum_{j=K}^k \bar{\alpha}^{k-j} \delta_j + \bar{\alpha}^{k-K} \sum_{j=1}^K \delta_j.$$

Careful examination of this expression yields

$$\sum_{k=0}^{\infty} [\theta_{k+1}]_+ \leq K \sum_{k=0}^K \delta_k + \frac{\bar{\alpha}^K}{1 - \bar{\alpha}} \left( [\theta_1]_+ + \sum_{k=K}^{\infty} \delta_k \right) < \infty. \quad (4.13)$$

Set  $w_k \triangleq \varphi_k - \sum_{j=0}^k [\theta_j]_+$ . Since  $\varphi_k \geq 0$  and  $\sum [\phi_j]_+ < \infty$ ,  $w_k$  is bounded from below.  $w_k$  is non-increasing, therefore we have it converges. Therefore  $\varphi_k$  converges for every  $x^* \in X^*$ . By invoking Opial's Lemma, statement (vi) is established.  $\square$

Theorem 4.1 does not apply to the FISTA-like parameter choices because for all of these choices  $\alpha_k \rightarrow 1$ . However the theorem does apply if we make the following modification. Replace the momentum parameter sequence  $\{\alpha_k\}_{k \in \mathbb{N}}$  with  $\min(\alpha_k, \bar{\alpha})$  where  $\bar{\alpha} < 1$ . This parameter choice satisfies the assumptions of the theorem, and  $\bar{\alpha}$  can be chosen arbitrarily close to 1. However the  $O(1/k^2)$  objective function convergence rate is no longer guaranteed once  $\alpha_k$  exceeds  $\bar{\alpha}$ .

In the following Corollary, we use (4.2) to determine explicit bounds on  $\sum_k \|\Delta_k\|^2$  which will be useful in the analysis of Problem SO.

**COROLLARY 4.2.** *Suppose that Assumption 1 holds. Assume  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and satisfies  $0 < \lambda_k \leq 1/L$  for all  $k$ ,  $\{\epsilon_k\}_{k \in \mathbb{N}}$  satisfies  $\epsilon_k \geq 0$  for all  $k$  and  $\sum_k \epsilon_k < \infty$ , there exists  $\bar{\alpha} \in [0, 1)$  such that  $\{\alpha_k\}_{k \in \mathbb{N}}$  satisfies  $0 \leq \alpha_k \leq \bar{\alpha}$  for all  $k$ . Then for the iterates of I-FBS,*

$$\sum_{k=1}^{\infty} \|\Delta_k\|^2 \leq \frac{2}{2L(1 - \bar{\alpha}) - 1} \left( F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 + \sum_{k=1}^{\infty} \epsilon_k \right). \quad (4.14)$$

If, in addition, there exists  $\underline{\alpha} \in [0, \bar{\alpha}]$  such that  $\alpha_k \geq \underline{\alpha}$  for all  $k$ , then

$$\sum_{k=1}^{\infty} \|\Delta_k\|^2 \leq \frac{2}{L^2 \underline{\alpha} (1 - \bar{\alpha})} \left( F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 + \sum_{k=1}^{\infty} \epsilon_k \right). \quad (4.15)$$

*Proof.* Inequality (4.2) implies

$$\begin{aligned} F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 + \sum_{k=1}^{\infty} \epsilon_k &\geq \sum_{k=1}^{\infty} \frac{2 + \lambda_k(\alpha_k - 1) - \alpha_k - \alpha_{k+1}}{2\lambda_k} \|\Delta_{k+1}\|^2 \\ &\geq \sum_{k=1}^{\infty} \frac{2 - 2\bar{\alpha} - \lambda_k}{2\lambda_k} \|\Delta_{k+1}\|^2 \end{aligned} \quad (4.16)$$

$$\geq \sum_{k=1}^{\infty} \frac{2L(1 - \bar{\alpha}) - 1}{2} \|\Delta_{k+1}\|^2 \quad (4.17)$$

which proves (4.14). To derive (4.16) we used the fact that  $0 \leq \alpha_k \leq \bar{\alpha}$ . To derive (4.17) we used the fact that  $\lambda L \leq 1$ .

Inequality (4.2) also implies

$$\begin{aligned} F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 + \sum_{k=1}^{\infty} \epsilon_k &\geq \sum_{k=1}^{\infty} \frac{L\alpha_k(1 - \alpha_k)}{2\lambda_k} \|\Delta_k\|^2 \\ &\geq \sum_{k=1}^{\infty} \frac{L^2\underline{\alpha}(1 - \bar{\alpha})}{2} \|\Delta_k\|^2 \end{aligned}$$

which proves (4.15).  $\square$

## 5. Convergence Analysis of I-FBS for Sparse Optimization.

**5.1. Finite Convergence Results.** We now turn our attention to Problem SO. The following theorem proves finite convergence to 0 for the components in  $D$ , and finite convergence to the correct sign for the components in  $E$  (recall the definitions of  $D$  and  $E$  in Section 2.4). Following the terminology of [30] we will refer to this as the “finite manifold identification period”. The manifold in this case is the half-space of vectors with support a subset of  $E$  and non-zero components with sign  $-h_i^*/\rho$ . This theorem generalizes the result of Theorem 4.5 in [17] from ISTA to I-FBS. For simplicity, we only consider the case where  $\epsilon_k$  is 0 for all  $k$ , meaning the proximal operator is computed exactly. Thus the results are stated for I-FBS not I-FBS- $\epsilon$ . Note that the proximal operator w.r.t. the  $\ell_1$  norm is relatively easy to compute as it is in separable closed form, thus we do not think it is worth considering I-FBS- $\epsilon$  in this case. In the next subsection, we consider the FISTA-like methods.

**THEOREM 5.1.** *Suppose that Assumption SO holds. Assume  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and satisfies  $0 < \lambda_k \leq 1/L$ , and there exists  $\underline{\alpha}, \bar{\alpha} \in [0, 1)$  such that  $\{\alpha_k\}_{k \in \mathbb{N}}$  satisfies  $\underline{\alpha} \leq \alpha_k \leq \bar{\alpha}$  for all  $k$ . Then, there exist constants  $K_D > 0$  and  $K_E > 0$  such that the iterates of I-FBS applied to Problem SO satisfy, for all  $k > K_E$ ,*

$$\text{sgn}(y_i^k - \lambda_k \nabla f(y^k)_i) = -\frac{h_i^*}{\rho}, \quad \forall i \in E, \quad (5.1)$$

and, for all  $k > K_D$

$$x_i^k = y_i^k = 0, \quad \forall i \in D. \quad (5.2)$$

Furthermore,  $K_E \leq \overline{K}_E$  and  $K_D \leq \overline{K}_D$ , where

$$\begin{aligned} \overline{K}_E \triangleq & \frac{1}{\rho^2 \lambda_1^2} \left[ \frac{2\overline{\alpha}(1 + \overline{\alpha}) \left( F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 \right)}{\underline{\alpha}(1 - \overline{\alpha})L^2} + \|x^1 - x^*\|^2 - \overline{\alpha}\|x^0 - x^*\|^2 \right] \\ & + \frac{\underline{\alpha}}{1 - \underline{\alpha}} \end{aligned} \quad (5.3)$$

and

$$\begin{aligned} \overline{K}_D \triangleq & \frac{1}{\omega^2 \lambda_1^2} \left[ \frac{2\overline{\alpha}(1 + \overline{\alpha}) \left( F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 \right)}{\underline{\alpha}(1 - \overline{\alpha})L^2} + \|x^1 - x^*\|^2 - \overline{\alpha}\|x^0 - x^*\|^2 \right] \\ & + \frac{\underline{\alpha}}{1 - \underline{\alpha}} + 2 \end{aligned} \quad (5.4)$$

for any  $x^* \in X^*$ .

*Proof.* Note that this parameter choice satisfies the requirements of Theorem 4.1 and Corollary 4.2. Furthermore, by assumption,  $X^*$  is non-empty and  $F^* \geq -\infty$ , thus all conclusions of Theorem 4.1 and Corollary 4.2 hold. Throughout the proof, fix an arbitrary  $x^* \in X^*$ .

**Proof of (5.1).** Fix a  $\lambda > 0$ . Recall from Theorem 2.1 there exists a vector  $h^*$  such that  $\nabla f(x^*) = h^*$  for all  $x^* \in X^*$ , and that  $\text{supp}(x^*) \subset E$ . For  $i \in \text{supp}(x^*)$ ,

$$0 \neq x_i^* = \text{sgn}(x_i^* - \lambda h_i^*) [|x_i^* - \lambda h_i^*| - \rho\lambda]_+. \quad (5.5)$$

Therefore  $|x_i^* - \lambda h_i^*| > \rho\lambda$  for all  $i \in \text{supp}(x^*)$ . On the other hand, if  $i \in E \setminus \text{supp}(x^*)$ , then

$$|x_i^* - \lambda h_i^*| = \lambda |h_i^*| = \rho\lambda.$$

Therefore

$$|x_i^* - \lambda h_i^*| \geq \rho\lambda, \quad \forall i \in E.$$

Looking at (5.5) it can be seen that

$$\text{sgn}(x_i^*) = \text{sgn}(x_i^* - \lambda h_i^*), \quad \forall i \in \text{supp}(x^*). \quad (5.6)$$

Note by Theorem 2.1, if  $i \in \text{supp}(x^*)$ , then  $\text{sgn}(x_i^*) = -h_i^*/\rho$ . Else if  $i \in E \setminus \text{supp}(x^*)$  then

$$\text{sgn}(x_i^* - \lambda h_i^*) = \text{sgn}(-\lambda h_i^*) = -\text{sgn}(h_i^*) = -\frac{h_i^*}{\rho}. \quad (5.7)$$

Combining (5.6) and (5.7) yields

$$\text{sgn}(x_i^* - \lambda h_i^*) = -h_i^*/\rho \quad \forall i \in E, \lambda > 0. \quad (5.8)$$

Let  $\nu_k = \rho\lambda_k$ . If

$$\text{sgn}(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) \neq \text{sgn}(x_i^* - \lambda_k h_i^*) = -h_i^*/\rho \quad \text{for some } i \in E, \quad (5.9)$$

then Lemma 2.3 implies

$$\begin{aligned}
|x_i^{k+1} - x_i^*|^2 &= |S_{\nu_k}(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) - S_{\nu_k}(x_i^* - \lambda_k h_i^*)|^2 \\
&\leq (|y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i - (x_i^* - \lambda_k h_i^*)| - \nu_k)^2 \\
&\leq |y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i - (x_i^* - \lambda_k h_i^*)|^2 - \nu_k^2
\end{aligned} \tag{5.10}$$

where (5.10) follows because

$$|(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) - (x_i^* - \lambda_k h_i^*)| > |(x_i^* - \lambda_k h_i^*)| \geq \nu_k > 0.$$

Using (5.10) we can say the following: Condition (5.9) implies that

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \sum_{j \neq i} |x_j^{k+1} - x^*|^2 + |x_i^{k+1} - x^*|^2 \\
&\leq \sum_{j \neq i} |y_j^{k+1} - \lambda_k \nabla f(y^{k+1})_j - (x_j^* - \lambda_k h_j^*)| \\
&\quad + |y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i - (x_i^* - \lambda_k h_i^*)|^2 - \nu_k^2
\end{aligned} \tag{5.11}$$

$$\begin{aligned}
&\leq \|y^{k+1} - \lambda_k \nabla f(y^{k+1}) - (x^* - \lambda_k h^*)\|^2 - \nu_k^2 \\
&\leq \|y^{k+1} - x^*\|^2 - \nu_k^2
\end{aligned} \tag{5.12}$$

$$\begin{aligned}
&= \|x^k + \alpha_k \Delta_k - x^*\|^2 - \nu_1^2 \\
&= \|x^k - x^*\|^2 + \alpha_k^2 \|\Delta_k\|^2 + 2\alpha_k \langle \Delta_k, x^k - x^* \rangle - \nu_1^2.
\end{aligned} \tag{5.13}$$

Inequality (5.11) follows from the element-wise nonexpansiveness of  $S_\nu$  along with (5.10). To deduce (5.12), we used Lemma 2.2. Finally, (5.13) follows because  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and therefore so is  $\{\nu_k\}$ .

Recall the definition of  $\varphi_k \triangleq \frac{1}{2} \|x^k - x^*\|^2$  and  $\theta_k \triangleq \varphi_k - \varphi_{k-1}$ . Now, moving  $\langle \Delta_k, \Delta_{k+1} \rangle$  to the other side of (4.7) reveals

$$\langle \Delta_k, x^k - x^* \rangle = \varphi_k - \varphi_{k-1} + \frac{1}{2} \|\Delta_k\|^2. \tag{5.14}$$

Substituting (5.14) into (5.13) yields

$$2(\varphi_{k+1} - \varphi_k) \leq 2\alpha_k(\varphi_k - \varphi_{k-1}) + \bar{\alpha}(1 + \bar{\alpha})\|\Delta_k\|^2 - \nu_1^2,$$

therefore

$$\theta_{k+1} \leq \alpha_k \theta_k + \frac{\bar{\alpha}(1 + \bar{\alpha})}{2} \|\Delta_k\|^2 - \frac{\nu_1^2}{2}.$$

Repeating the arguments that led to (4.13), we can say the following: if (5.9) is true then

$$\theta_{k+1} \leq \bar{\alpha}^k \theta_1 + \frac{\bar{\alpha}(1 + \bar{\alpha})}{2} \sum_{j=1}^k \bar{\alpha}^{k-j} \|\Delta_j\|^2 - \frac{\nu_1^2}{2} \sum_{j=1}^k \bar{\alpha}^{k-j}. \tag{5.15}$$

Therefore, for  $M \in \mathbb{Z}_+$ , if (5.9) holds at iteration  $M$ , then

$$\begin{aligned} \varphi_M - \varphi_0 &= \sum_{k=1}^M \theta_k \\ &\leq \frac{\theta_1(1 - \bar{\alpha}^M)}{1 - \bar{\alpha}} + \frac{\bar{\alpha}(1 + \bar{\alpha})}{2(1 - \bar{\alpha})} \sum_{k=1}^M \|\Delta_k\|^2 - \frac{\nu^2}{2} \sum_{k=1}^M \sum_{j=0}^{k-1} \underline{\alpha}^j \\ &\leq \frac{\theta_1}{1 - \bar{\alpha}} + \frac{\bar{\alpha}(1 + \bar{\alpha})}{2(1 - \bar{\alpha})} \sum_{k=1}^M \|\Delta_k\|^2 - \frac{\nu^2}{2} \left( \frac{M}{1 - \underline{\alpha}} - \frac{\underline{\alpha}}{(1 - \underline{\alpha})^2} \right). \end{aligned} \quad (5.16)$$

To derive (5.16) we lower bounded the coefficient of  $-\frac{\nu^2}{2}$ . Since  $\varphi_k \geq 0$ , if (5.9) is true at iteration  $k$  then

$$\begin{aligned} k &\leq \frac{2(1 - \underline{\alpha})}{\nu_1^2} \left[ \frac{\bar{\alpha}(1 + \bar{\alpha})}{2(1 - \bar{\alpha})} \sum_{k=1}^{\infty} \|\Delta_k\|^2 + \frac{\|x^0 - x^*\|^2}{2} + \frac{\theta_1}{1 - \bar{\alpha}} \right] + \frac{\underline{\alpha}}{1 - \underline{\alpha}} \\ &\leq \frac{1}{\nu_1^2} \left[ \frac{2\bar{\alpha}(1 + \bar{\alpha}) \left( F(x^1) - F^* + \frac{\alpha_1}{2\lambda_1} \|\Delta_1\|^2 \right)}{\underline{\alpha}(1 - \bar{\alpha})L^2} + \|x^1 - x^*\|^2 - \bar{\alpha}\|x^0 - x^*\|^2 \right] \\ &\quad + \frac{\underline{\alpha}}{1 - \underline{\alpha}} \end{aligned} \quad (5.17)$$

To derive (5.18) we used the upper bound on  $\sum_k \|\Delta_k\|^2$  in (4.15) from Corollary 4.2. This upper bound is tighter than the other upper bound for  $\sum_k \|\Delta_k\|^2$  given in (4.14), so long as  $L > 2/\underline{\alpha}$ .

**Proof of (5.2).** Recall the definition of  $\omega$  and note that

$$\lambda_k \omega = \min\{\nu_k - |x_i^* - \lambda_k h_i^*| : i \in D\} > 0 \quad (5.19)$$

Consider  $i \in D$  (which implies  $i \notin \text{supp}(x^*)$ ). If  $x_i^{k+1} \neq 0$ , then Lemma 2.3 implies

$$\begin{aligned} |x_i^{k+1}|^2 &= |S_{\nu_k}(y_i^{k+1} - \lambda_k \nabla f(y_i^{k+1})) - S_{\nu_k}(x_i^* - \lambda_k h_i^*)|^2 \\ &\leq [|(y_i^{k+1} - \lambda_k \nabla f(y_i^{k+1})) - (x_i^* - \lambda_k h_i^*)| - (\nu_k - |x_i^* - \lambda_k h_i^*|)]^2 \\ &\leq |(y_i^{k+1} - \lambda_k \nabla f(y_i^{k+1})) - (x_i^* - \lambda_k h_i^*)|^2 - (\nu_k - |x_i^* - \lambda_k h_i^*|)^2 \end{aligned} \quad (5.20)$$

$$\leq |(y_i^{k+1} - \lambda_k \nabla f(y_i^{k+1})) - (x_i^* - \lambda_k h_i^*)|^2 - \omega^2 \lambda_k^2. \quad (5.21)$$

To derive (5.20) we used the fact that

$$|(y_i^{k+1} - \lambda_k \nabla f(y_i^{k+1})) - \lambda_k h_i^*| > \nu_k - \lambda_k |h_i^*|.$$

To derive (5.21) we used (5.19). Repeating the arguments used to prove (5.13) we can say the following. If there exists  $i \in D$ , such that  $x_i^{k+1} \neq 0$ , then  $k \leq \bar{K}_D - 2$ , with  $\bar{K}_D$  defined in (5.4). Therefore  $|x_i^k| = 0$  for all  $k > \bar{K}_D - 2$ . Since  $y_i^{k+1} = x_i^k + \alpha_k(x_i^k - x_i^{k-1})$ ,  $y_i^k = 0$  for all  $i \in D$  and  $k > \bar{K}_D - 2 + 2 = \bar{K}_D$ , which proves (5.2).  $\square$

Note that if  $x_i^{k+1} \neq 0$  then  $\text{sgn}(x_i^{k+1}) = y_i^k - \lambda_k \nabla f(y_i^k)_i$ , thus (5.1) implies convergence in sign. We can recover the result by Hale et al. for ISTA (Theorem 4.5 [17]). To see this, consider (5.17) with  $\bar{\alpha} = \underline{\alpha} = 0$  and then use the upper bound on  $\sum_k \|\Delta_k\|^2$  given in (4.14) of Corollary 4.2. Note that we defined the constant  $\omega$  in a slightly differently way to Hale et al.



**5.2. Finite Convergence of FISTA.** We can prove convergence to the optimal manifold in a finite number of iterations under more general conditions than required in Theorem 5.1, however without explicit bounds on the number of iterations. A corollary of the following theorem is that the FISTA-like choices proposed by Beck and Teboulle [16] and Chambolle and Dossal [27] achieve finite manifold identification.

**THEOREM 5.2.** *Suppose that Assumption SO holds. Assume  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and satisfies  $0 < \lambda_k < 2/L$  for all  $k$ , and there exists  $\bar{\alpha} \geq 0$  such that  $\{\alpha_k\}_{k \in \mathbb{N}}$  satisfies  $0 \leq \alpha_k \leq \bar{\alpha}$ , for all  $k$ . If, for the iterates of I-FBS applied to Problem SO, it is true that  $\sum_{k=1}^{\infty} \|\Delta_k\|^2 < \infty$  and  $\|x^k - x^*\|$  is bounded for some  $x^* \in X^*$  and for all  $k$ , then there exists a constant  $K > 0$  such that for the iterates of I-FBS (5.1) and (5.2) hold for all  $k > K$ .*

*Proof.* Inequality (5.13) and the equivalent recursion for when (5.9) holds can be proved in exactly the same way. However, we cannot rely on  $\bar{\alpha} < 1$ , so we have to modify the proof from that point onwards. Once again, fix  $x^* \in X^*$ . Rewriting (5.13), we can say that (5.9) implies that

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 + \bar{\alpha}^2 \|\Delta_k\|^2 + 2\bar{\alpha} \langle \Delta_k, x^k - x^* \rangle - \nu_1^2 \\ &\leq \|x^k - x^*\|^2 + \bar{\alpha}^2 \|\Delta_k\|^2 + 2\bar{\alpha} \|\Delta_k\| \|x^k - x^*\| - \nu_1^2 \end{aligned} \quad (5.22)$$

$$\leq \|x^k - x^*\|^2 + \bar{\alpha}^2 \|\Delta_k\|^2 + 2\bar{\alpha} M_1 \|\Delta_k\| - \nu_1^2 \quad (5.23)$$

where we used the Cauchy-Schwarz inequality to get (5.22). To derive (5.23) we used the assumption that there exists  $M_1 > 0$  such that  $\|x^k - x^*\| < M_1$ . Also by assumption, there exists  $M_2 > 0$  such that

$$\sum_{k=1}^{\infty} \|\Delta_k\|^2 < M_2. \quad (5.24)$$

Now, by Jensen's inequality

$$\begin{aligned} \sum_{i=0}^k \|\Delta_i\| &\leq \sqrt{k \sum_{i=0}^k \|\Delta_i\|^2} \\ &\leq \sqrt{M_2 k}. \end{aligned} \quad (5.25)$$

Substituting (5.24) and (5.25) into (5.23) yields the following: If (5.9) is true then

$$\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 + \bar{\alpha}^2 M_2 + 2\bar{\alpha} M_1 \sqrt{M_2 k} - k\nu_1^2 \quad (5.26)$$

The r.h.s. of (5.26) can be non-negative for only a finite number of iterations, which proves (5.1).

For (5.2), the recursion is

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 + \bar{\alpha}^2 \|\Delta_k\|^2 + 2\bar{\alpha} \langle \Delta_k, x^k - x^* \rangle - \omega^2 \lambda_1^2$$

and the reasoning is the same from this point onwards.  $\square$

The classical FISTA parameter choice in (3.1) due to Beck and Teboulle [16], along with others which guarantee the  $O(1/k^2)$  rate provided by Tseng [26], satisfy the assumptions of Theorem 5.2, so long as  $F$  is coercive (or equivalently has bounded level-sets). The first condition,  $\sum_k \|\Delta_k\|^2 < \infty$  can be shown by considering the

following facts. The sequence  $b_k$  defined on page 196 of [16] is bounded, which implies the sequence  $u_k$  defined on page 194 of [16] is also bounded. If  $F$  is coercive, then  $x^k$  is bounded, since  $F(x^k) \rightarrow F^*$ . This implies  $\|x^k - x^*\|$  is bounded for some  $x^* \in X^*$ . It also implies  $t_k \Delta_k$  is bounded and since  $t_k = O(k)$ ,  $\|\Delta_k\| = O(1/k)$ , and the result follows<sup>2</sup>.

The parameter choice (3.2) due to Chambolle and Dossal [27] satisfies the assumptions of this theorem, even when  $F$  is not assumed to be coercive.  $\sum_k \|\Delta_k\|^2$  is finite by Corollary 2 of [27] and  $\|x^k - x^*\|$  is shown to be bounded for all  $x^* \in X^*$  in the proof of Theorem 3 of [27]. (In fact Chambolle and Dossal proved that  $\sum_k k \|\Delta_k\|^2$  is finite for their parameter choice.)

**5.3. Finite Reduction to Local Minimization.** Theorems 5.1 and 5.2 allow us to characterize the behavior of I-FBS (including the FISTA-like choices) after a finite manifold identification period. In the following corollary, we show that after a finite number of iterations, I-FBS reduces to minimizing a smooth function over  $E$  subject to an orthant constraint. The following corollary generalizes the result of Corollary 4.6 in [17] from ISTA to I-FBS.

**COROLLARY 5.3.** *Suppose that Assumption SO holds. Assume  $\{\lambda_k\}_{k \in \mathbb{N}}$  is non-decreasing and satisfies  $0 < \lambda_k \leq 1/L$ , and there exists  $\underline{\alpha}, \bar{\alpha} \in [0, 1)$  such that  $\underline{\alpha} \leq \bar{\alpha}$  and  $\{\alpha_k\}_{k \in \mathbb{N}}$  satisfies  $\underline{\alpha} \leq \alpha_k \leq \bar{\alpha}$  for all  $k$ . Then, after finitely many iterations, the iterates of I-FBS applied to Problem SO become equivalent to the iterates of I-FBS applied to minimizing  $\phi : \mathbb{R}^{|E|} \rightarrow \mathbb{R}$ , where*

$$\phi(x_E) \triangleq -(h_E^*)^T x_E + f((x_E, 0)), \quad (5.27)$$

constrained to the orthant  $O_E$ , where

$$O_E \triangleq \{x_E \in \mathbb{R}^{|E|} : -\text{sgn}(h_i^*) x_i \geq 0, \forall i \in E\}. \quad (5.28)$$

Specifically, there exists  $K > 0$  such that for all  $k > K$ ,

$$y_E^{k+1} = x_E^k + \alpha_k(x_E^k - x_E^{k-1}) \quad (5.29)$$

$$x_E^{k+1} = P_{O_E}(y_E^{k+1} - \lambda_k \nabla \phi(y_E^{k+1})), \quad (5.30)$$

$x_D^k = y_D^k = 0$ , and  $F(x^k) = \phi(x_E^k)$ . Furthermore  $K \leq \max\{\bar{K}_D, \bar{K}_E\}$ , with  $\bar{K}_D$  and  $\bar{K}_E$  defined in (5.3) and (5.4).

Alternatively, if the conditions of Theorem 5.2 hold, then there exists  $K' > 0$  such that (5.29)-(5.30) hold,  $x_D^k = y_D^k = 0$ , and  $F(x^k) = \phi(x_E^k)$ , for all  $k > K'$ .

*Proof.* From Theorem 5.1, there exists a  $K$  such that for all  $k > K$ , (5.1) and (5.2) hold and  $K \leq \max\{\bar{K}_D, \bar{K}_E\}$ . Take  $k > K$ . Since  $x_i^k = 0$  for all  $i \in D$  it suffices to consider  $i \in E$ . For  $i \in E$ ,  $k > K$ , using (5.1) we have

$$x_i^k \geq 0 \text{ if } \text{sgn}(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) = 1 \text{ equivalently } h_i^* < 0$$

and

$$x_i^k \leq 0 \text{ if } \text{sgn}(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) = -1 \text{ equivalently } h_i^* > 0.$$

Therefore for any  $i \in E$ ,  $-h_i^* x_i^k \geq 0$ , thus  $x_E^k \in O_E$  for all  $k > K$ . Next note that, for all  $i \in E$ ,  $-h_i^* x_i^k = \rho |x_i|$ . Therefore for  $k > K$ ,  $-(h^*)^T x_E^k = \rho \|x^k\|_1$ , thus  $F(x^k) = \phi(x_E^k)$ .

<sup>2</sup>We thank Antonin Chambolle and Charles Dossal for pointing this out to us.

Now for  $i \in E$ ,  $k > K$ , we calculate the quantity

$$\begin{aligned} z_i^{k+1} &\triangleq y_i^{k+1} - \lambda_k \nabla \phi(y_E^{k+1})_i \\ &= y_i^{k+1} - \lambda_k (-h_i^* + \nabla f(y^{k+1})_i) \\ &= y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i + \rho \lambda_k \left( \frac{h_i^*}{\rho} \right) \\ &= \text{sgn}(y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) (|y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i| - \rho \lambda_k). \end{aligned}$$

Therefore, for  $i \in E$ ,  $k > K$ ,

$$\begin{aligned} y_i^{k+1} &= x_i^k + \alpha_k (x_i^k - x_i^{k-1}), \\ x_i^{k+1} &= S_{\rho \lambda_k} (y_i^{k+1} - \lambda_k \nabla f(y^{k+1})_i) = \begin{cases} z_i^{k+1} : -h_i^* z_i^{k+1} \geq 0 \\ 0 : \text{else} \end{cases} \end{aligned}$$

Equivalently, for  $k > K$ ,

$$\begin{aligned} y_E^{k+1} &= x_E^k + \alpha_k (x_E^k - x_E^{k-1}), \\ x_E^{k+1} &= P_{O_E} (y_E^{k+1} - \lambda_k \nabla \phi(y_E^{k+1})). \end{aligned}$$

Due to Theorem 5.2, the same arguments hold for parameter choices such that  $\sum_k \|\Delta_k\|^2$  is finite,  $\|x^k - x^*\|$  is bounded for all  $k$  and some  $x^* \in X^*$ , and  $\alpha_k$  is bounded. However there is no explicit upper bound on  $K$ .  $\square$

In principle one could switch to minimizing  $\phi$  directly once the algorithm has reduced to (5.29)-(5.30). This would allow for a larger step-size, since the Lipschitz constant of  $\nabla \phi$  is less than  $L$ . However it is not possible to know with certainty that the algorithm has transitioned to the form (5.29)-(3.2) unless the number of iterations exceeds the upper bound  $\max\{\overline{K}_D, \overline{K}_E\}$ , although we discuss some heuristics for identifying this transition in Section 5.5. The main drawback of this strategy is that once it switches to minimizing  $\phi$  directly the support of  $x^k$  is fixed. Therefore any mismatch between  $\text{supp}(x^k)$  and  $\text{supp}(x^*)$  is not identified and the algorithm will not necessarily converge to an optimal point. In the next section we discuss a method which uses the optimal momentum for minimizing  $\phi$  yet continues to use a smaller step-size and is therefore guaranteed to converge to a minimizer.

**5.4. A Simple Locally Optimal Parameter Choice for I-FBS.** The analysis of the previous three sections shows that, after a finite number of iterations, I-FBS (subject to parameter conditions) reduces to minimizing the function  $\phi$  subject to an orthant constraint. Even though  $f$  is not assumed to be strongly convex,  $\phi$  might be. If this function is strongly convex, the asymptotic rate of convergence can be determined by the worst-case condition number of the Hessian. Throughout this section let  $H_{EE}(v)$  be the Hessian of  $\phi$  evaluated at  $v$ . In terms of strategies for choosing  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$ , one approach is to choose them to obtain the best iteration complexity for minimizing  $\phi$ . In the following Corollary, we provide a simple fixed choice which does this and thus optimizes the asymptotic iteration complexity.

**COROLLARY 5.4.** *Suppose that Assumption SO holds, and  $\phi$  is strongly convex. Let  $x^*$  be the unique minimizer of Problem SO and  $l_E$  be the strong convexity parameter of  $\phi$ . If  $\lambda \in (0, 1/L]$ ,*

$$\lambda_k = \lambda \text{ and } \alpha_k = \frac{1 - \sqrt{l_E \lambda}}{1 + \sqrt{l_E \lambda}} \quad \forall k \in \mathbb{Z}_+, \quad (5.31)$$

then the iterates  $\{x^k\}_{k \in \mathbb{N}}$  of I-FBS converge to  $x^*$  linearly and  $F(x^k)$  converges to  $F^*$  linearly. Indeed

$$F(x^k) - F^* = O\left(\left(1 - \sqrt{l_E \lambda}\right)^k\right). \quad (5.32)$$

*Proof.* The analysis of the previous sections shows that, for the given choice of  $\{\alpha_k\}_{k \in \mathbb{N}}$  and  $\{\lambda_k\}_{k \in \mathbb{N}}$  there exists a  $K$  such that, for all  $k > K$ , (5.29) and (5.30) hold, and  $F(x^k) = \phi(x_E^k)$ . Thus for  $k > K$  the algorithm is equivalent to Nesterov's method (Section 2.2 [25]) applied to the strongly convex function  $\phi$  subject to the orthant constraint  $O_E$ . Therefore we apply Theorem 2.2.1 of [25] with the fixed parameter choice (constant scheme 3, discussed on p. 76 of [25]). The only difference compared to Theorem 2.2.1 is that we allow for step-sizes other than  $1/L_E$ , where  $L_E$  is the Lipschitz constant of  $\nabla \phi$ . Note that  $L \geq L_E$ . This minor change is discussed on p. 72. of [25]. Setting  $\lambda = 1/L$  (the maximum allowed step-size) gives:

$$F(x^k) - F^* = O\left(\left(1 - \sqrt{\frac{l_E}{L}}\right)^k\right). \quad (5.33)$$

Another minor issue to note is that the minimization is constrained to the simple convex set  $O_E$ . This does not effect the convergence of Nesterov's method, as discussed in Constant Step Scheme (2.2.17) of [25].

By the strong convexity of  $\phi$ , the sequence  $\{x^k\}_{k \in \mathbb{N}}$  also achieves linear convergence with the same iteration complexity.  $\square$

The iteration complexity with this parameter choice is

$$\Omega\left(\sqrt{\frac{L}{l_E}} \log\left(\frac{1}{\epsilon}\right)\right). \quad (5.34)$$

This is the best asymptotic iteration complexity that can be achieved by I-FBS using this step-size [25]. Therefore we will refer to it as the locally optimal choice. Indeed it is better than the iteration complexity of ISTA [17] (which corresponds to I-FBS with  $\alpha_k$  equal to 0) which is

$$\Omega\left(\frac{L}{l_E} \log\left(\frac{1}{\epsilon}\right)\right).$$

We will see in the next section that (5.34) is better than the iteration complexity achieved by the FISTA-like choices of [16], [26] and [27].

In practice the optimal momentum is not known a priori as it depends on the smallest eigenvalue of  $H_{EE}$ . The momentum could be estimated periodically based on the smallest eigenvalue of the Hessian corresponding to the current support set, or adapted based on the behavior of  $F(x^k)$  (see Section 5.6 and Section 6).

The authors of [34] proposed an adaptive momentum restart scheme for Nesterov's method in the case of smooth optimization (i.e.  $g(x) = 0$ ). Corollary 5.3 implies that the scheme can also be used for I-FBS applied to Problem SO. This follows because I-FBS with parameter choice (3.1) reduces to minimizing a smooth function  $\phi$  after a finite number of iterations, after which the momentum restart scheme can be used. Referring to the analysis of [34], it can be shown that the method will have the same iteration complexity as given in (5.34).

Local linear convergence can also be proved when the local function  $\phi$  is not strongly convex, but the limit point of the iterations obeys the “strict-complementarity” condition:  $E = \text{supp}(x^*)$ . Furthermore the Hessian matrix of  $\phi$  must be invariant in a region containing the limit. In the following corollary, let  $x^* = \lim_{k \rightarrow \infty} x^k$ , which is in  $X^*$  by Theorem 4.1.

**COROLLARY 5.5.** *Suppose Assumption SO holds and  $E = \text{supp}(x^*)$ , where  $\lim_{k \rightarrow \infty} x^k = x^* \in X^*$ . Let  $H_{EE}(x)$  be the Hessian of the function  $\phi$  defined in (5.27). Let  $\hat{l}_E$  be the smallest non-zero eigenvalue of  $H_{EE}(x_E^*)$ . Assume the range space of  $H_{EE}$  is invariant in some neighborhood  $N^*$  around  $x^*$ . If all eigenvalues of  $H_{EE}(x_E^*)$  are zero,  $x^k = x^*$  after a finite number of iterations, for any choice of  $\{\lambda_k, \alpha_k\}$  satisfying the conditions of Theorem 5.1 or Theorem 5.2. If  $\hat{l}_E > 0$ ,  $\lambda \in (0, 1/L]$ ,*

$$\lambda_k = \lambda \text{ and } \alpha_k = \frac{1 - \sqrt{\hat{l}_E \lambda}}{1 + \sqrt{\hat{l}_E \lambda}} \quad \forall k \in \mathbb{Z}_+, \quad (5.35)$$

*then the iterates  $x^k$  of I-FBS converge to  $x^*$  linearly and  $F(x^k)$  converges to  $F^*$  linearly. Indeed*

$$F(x^k) - F^* = O\left(\left(1 - \sqrt{\hat{l}_E \lambda}\right)^k\right).$$

*Proof.*

The proof proceeds almost identically to Theorem 4.11 of [17]. Note that if  $x^k \rightarrow x^*$  than  $y^k \rightarrow x^*$ . Now Lemma 5.3 of [17] can be directly applied to I-FBS to say that, after a finite number of iterations,

$$x_i^{k+1} = y_i^{k+1} - \lambda_k (\nabla f(y^{k+1})_i - h_i^*) \quad \forall i \in \text{supp}(x^*). \quad (5.36)$$

Assume  $k$  is large enough that (5.36) holds,  $x^k \in N^*$ , and  $k > \max\{\bar{K}_D, \bar{K}_E\}$ . Since  $x_i^k$  is 0 for all  $i \in D$ , it suffices to consider  $i \in E = \text{supp}(x^*)$ . Recall that  $H(x)$  is the Hessian of  $f$  evaluated at  $x$ . Now, let  $\bar{H}^k$  be defined as

$$\bar{H}^k \triangleq \int_0^1 H(x^* + t(x^k - x^*)) dt$$

By assumption the range spaces of  $\bar{H}^k$  are now invariant over  $k$ . For a matrix  $W$ , let  $W_{EE}$  be the  $|E| \times |E|$  submatrix of  $W$  with row and column indices given by  $E$ . Let  $P$  be the orthogonal projection onto the range space of  $H_{EE}$ . Since  $E = \text{supp}(x^*)$ , equation (5.36) can be used to claim that

$$x_E^{k+1} = y_E^{k+1} - \lambda (\nabla f(y^{k+1}) - h^*)_E = y_E^{k+1} - \lambda \bar{H}_{EE}^k (y_E^{k+1} - x_E^*). \quad (5.37)$$

This follows from the mean value theorem. At each iteration the term  $-\lambda \bar{H}_{EE}^k (y_E^{k+1} - x_E^*)$  stays in the range space of  $H_{EE}$ , which implies that the null-space components of the iterates have already converged. In other words, for  $k$  sufficiently large,

$$(I - P)(x_E^k - x_E^*) = 0.$$

If  $\hat{l}_E > 0$ , it suffices to consider the convergence of  $\{Px^k\}$ , that is, consider the component in the range space of  $H_{EE}$ . Since  $\phi$  restricted to the range space of  $H_{EE}$  is strongly convex, we now simply repeat the arguments of Corollary 5.3 and the result follows.  $\square$

**5.5. “Underdamped” I-FBS.** In [34], the behavior of the FISTA-like methods when applied to strongly convex functions was investigated. It was shown that for such functions if  $\alpha_k \approx 1$ , the algorithm moves into what is known as an “underdamped regime”, which leads to oscillations in the objective function at a predictable frequency, and a sub-optimal iteration complexity. The results of the preceding sections allow us to extend this analysis to Problem SO, despite it being nonsmooth and not strictly convex.

The analysis of [34] revealed that if  $\alpha_k \approx 1$ , the trace of the objective function values will oscillate with a frequency proportional to  $1/\sqrt{\kappa}$  where  $\kappa$  is the condition number of the Hessian at the minimum. The iteration complexity in the high-momentum regime with step-size  $\lambda = 1/L$  is

$$\Omega\left(\kappa \log\left(\frac{1}{\epsilon}\right)\right)$$

More precisely, the behavior  $F(x^k) \approx C(1 - \frac{1}{\kappa})^k \cos^2(k/\sqrt{\kappa})$  is observed. Now Corollary 5.3 shows that after a finite number of iterations, FISTA (with parametric constraints) reduces to minimizing  $\phi$  (defined in (5.27)) subject to an orthant constraint. Therefore we can apply the analysis of [34] once the algorithm is in this regime. Thus if  $l_E > 0$ , FISTA obeys the conditions of Theorem 5.1 or 5.2, and  $\alpha_k \rightarrow 1$ , then the iteration complexity will be [34]

$$\Omega\left(\frac{L}{l_E} \log\left(\frac{1}{\epsilon}\right)\right),$$

which is worse than the iteration complexity achieved with the locally optimal choice given in Corollary 5.4. The trace of the objective function will exhibit oscillations with period  $\sqrt{l_E/L}$ . This result directly applies to the FISTA-like choices of [16], [26] and [27]. The result also applies when  $l_E = 0$  under the strict-complementarity condition. In this case replace  $l_E$  with  $\hat{l}_E$  defined in Corollary 5.5.

**5.6. An Adaptive Modification.** In numerical experiments we have noticed that it can take many iterations for the optimal manifold to be identified by I-FBS. This means that one of the FISTA-like choices can outperform I-FBS with our locally optimal choice (5.31) before the optimal manifold is identified. This is because the FISTA-like choices guarantee  $O(1/k^2)$  convergence during this phase whereas the locally optimal choice does not have a guaranteed rate until the optimal manifold is identified (although an  $O(1/k)$  rate can probably be established following the analysis of [27], however this is beyond the scope this paper). On the other hand the analysis of the previous section showed that the FISTA-like choices have poor performance once the algorithm is in the optimal manifold. In summary the FISTA-like choices have excellent global properties but poor local properties.

In light of this we propose the following adaptive heuristic. We use the condition,  $F(x^k) > F(x^{k-1})$ , as an indication the algorithm is at least approximately operating in the optimal manifold. This is because with a FISTA-like choice the algorithm will eventually converge to the optimal manifold and then the function values will start to oscillate. So the adaptive modification is the following. Use Beck and Teboulle’s parameter choice of (3.1) until  $F(x^k) > F(x^{k-1})$ . For all iterations after, use the locally optimal momentum given in (5.31). We call this scheme FISTA-AdOPT. See Experiment 1 for empirical results. It is worth mentioning that it is better to use the condition  $(y^{k+1} - x^{k+1})^T(x^{k+1} - x^k) > 0$  rather than  $F(x^k) > F(x^{k-1})$ . It was shown

in [34] that the two conditions are equivalent however the first avoids computation of  $F$ .

The iterates of FISTA-AdOPT are guaranteed to converge to a solution and, until  $F(x^k) > F(x^{k-1})$  occurs, the convergence in the objective function is  $O(1/k^2)$ . Furthermore, the asymptotic convergence rate is optimal and given by (5.33). The main drawback of the scheme is that it might switch to the locally optimal parameter choice before the iterates are confined to the optimal manifold. Another drawback is that the locally optimal momentum parameter must be estimated, which involves computing the largest and smallest singular values of  $A$  confined to the current support set. This computation is nontrivial when the support set is large. A better alternative is to use the adaptive restart scheme proposed in [34]. Thanks to our analysis, this method is guaranteed to achieve the same asymptotic iteration complexity as the locally optimal parameter choice, but will also achieve the  $O(1/k^2)$  performance in the transient regime prior to manifold identification (See the remarks after Corollary 5.4). The practical performance of the momentum restart scheme and FISTA-AdOPT are compared in the next section.

**5.7. The Splitting Inertial Proximal Method.** In [35], Moudafi and Oliny introduced the Splitting Inertial Proximal Method (SIPM):

$$x^{k+1} \in -\lambda_k \partial_{\epsilon_k} g(x^{k+1}) + x^k - \lambda_k \nabla f(x^k) + \alpha_k (x^k - x^{k-1}). \quad (5.38)$$

In fact they introduced it for the more general monotone inclusion problem. The method is a direct generalization of Polyak's heavy-ball with friction method to problems involving the sum of two functions. It differs from I-FBS in that the gradient w.r.t.  $f$  is computed at  $x^k$  rather than the extrapolated point  $x^k + \alpha_k (x^k - x^{k-1})$ . Our analysis of I-FBS in the case of sparse optimization can be extended easily to Moudafi and Oliny's method under the condition that  $\sum_k \|\Delta_k\|^2$  is finite, for which sufficient conditions were established in [35].

**THEOREM 5.6.** *Suppose that Assumption SO holds. Assume  $\epsilon_k$  is 0 for all  $k \in \mathbb{N}$ ,  $0 < \lambda_k < 2/L$  for all  $k$ , and there exists  $\bar{\alpha} \geq 0$  such that  $0 \leq \alpha_k \leq \bar{\alpha}$  for all  $k$ . If  $\sum_k \|\Delta_k\|^2 < \infty$ , and  $\|x^k - x^*\|$  is bounded for all  $x^* \in X^*$ , then there exists a constant  $K > 0$  such that, for all  $k > K$  the iterates of Algorithm (5.38) applied to Problem SO satisfy*

$$\text{sgn}(x_i^k - \lambda_k \nabla f(x^k)_i + \alpha_k (x_i^k - x_i^{k-1})) = -\frac{h_i^*}{\rho}, \quad \forall i \in E,$$

and

$$x_i^k = 0, \quad \forall i \in D.$$

*Proof.* The proof follows in a similar way to Theorems 5.1 and 5.2. Equation (5.22) is proved by following similar arguments as in the proof of Theorem 5.1. We include the salient differences.

Recall that  $\nu_k = \rho \lambda_k$ . If

$$\begin{aligned} \text{sgn}(x_i^k - \lambda_k \nabla f(x^k)_i + \alpha_k (x_i^k - x_i^{k-1})) &\neq \text{sgn}(x_i^* - \lambda_k h_i^*) = -h_i^*/\rho \\ &\text{for some } i \in E, \end{aligned} \quad (5.39)$$

then,

$$\begin{aligned} |x_i^{k+1} - x_i^*|^2 &= |S_{\nu_k}(x_i^k - \lambda_k \nabla f(x^k)_i + \alpha_k(x_i^k - x_i^{k-1})) - S_{\nu_k}(x_i^* - \lambda_k h_i^*)|^2 \\ &\leq |x_i^k - \lambda_k \nabla f(x^k)_i + \alpha_k(x_i^k - x_i^{k-1}) - (x_i^* - \lambda_k h_i^*)|^2 - \nu_k^2 \end{aligned} \quad (5.40)$$

where (5.40) follows for the same reasons as given for (5.10). Continuing to follow the proof of Theorem 5.1, we say the following: if (5.39) holds, then

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \sum_{j \neq i} |x_j^{k+1} - x^*|^2 + |x_i^{k+1} - x^*|^2 \\ &\leq \sum_{j \neq i} |x_j^k - \lambda_k \nabla f(x^k)_j + \alpha_k(x_j^k - x_j^{k-1}) - (x_j^* - \lambda_k h_j^*)|^2 \\ &\quad + |x_i^k - \lambda_k \nabla f(x^k)_i + \alpha_k(x_i^k - x_i^{k-1}) - (x_i^* - \lambda_k h_i^*)|^2 \\ &\quad - \nu_k^2 \end{aligned} \quad (5.41)$$

$$\begin{aligned} &\leq \|x^k - \lambda_k \nabla f(x^k) + \alpha_k \Delta_k - (x^* - \lambda_k h^*)\|^2 - \nu_1^2 \\ &\leq \|x^k - x^*\|^2 + \bar{\alpha}^2 \|\Delta_k\|^2 + 2\bar{\alpha} \|x^k - x^*\| \|\Delta_k\| \end{aligned} \quad (5.42)$$

Equation (5.41) follows from the element-wise nonexpansiveness of  $S_{\nu}$  and (5.40), and (5.42) follows from the nonexpansiveness of  $I - \lambda \nabla f$ , by application of Cauchy-Schwartz, and by substituting the upper bound  $\bar{\alpha}$  for  $\alpha_k$ . Equation (5.42) is identical to (5.22). From this point on, the proof is identical to Theorem 5.2. As before we cannot explicitly bound the number of iterations unless we know an upper bound for  $\sum_k \Delta_k^2$ .  $\square$

Moudafi and Oliny provided a condition on the parameters under which

$$\sum_k \|\Delta_k\|^2 < \infty,$$

and  $\|x^k - x^*\|$  is bounded for all  $x^* \in X^*$ . Choose the sequence  $\{\alpha_k\}_{k \in \mathbb{N}}$  to be non-decreasing, the constant  $\bar{\alpha} < 1/3$ , and the sequence  $\{\lambda_k\}_{k \in \mathbb{N}}$  to be bounded away from  $2/L$ . They showed that this implies  $\|x^k - x^*\|$  is bounded for all  $k$  and  $x^* \in X^*$  and

$$\sum_k \alpha_k \|\Delta_k\|^2 < \infty.$$

Now using the fact that  $\alpha_k$  is less than  $1/3$ ,  $\sum_k \|\Delta_k\|^2 < \infty$ . If  $\alpha_k$  is zero for all  $k$  the theorem follows from [17] since the algorithm reduces to ISTA.

We have proved that finite convergence in sign on  $E$  and to 0 on  $D$  holds for SIPM applied to Problem SO. An analogous result to Corollary 5.3 can now be shown. After a finite number of iterations, SIPM reduces to HBF projected onto a quadrant. Because of its similarity to Corollary 5.3, the proof is omitted.

**COROLLARY 5.7.** *Assume  $0 < \lambda_k < 2/L$  for all  $k$ , and there exists  $\bar{\alpha} \geq 0$  such that  $0 \leq \alpha_k \leq \bar{\alpha}$  for all  $k$ . If  $\sum_k \|\Delta_k\|^2 < \infty$ , and  $\|x^k - x^*\|$  is bounded for all  $x^* \in X^*$ , then there exists a  $K > 0$  such that for  $k > K$ , the iterates of SIPM satisfy*

$$\begin{aligned} x_E^{k+1} &= P_{O_E}(x_E^k - \lambda_k \nabla \phi(x_E^k) + \alpha_k(x_E^k - x_E^{k-1})) \\ x_D^k &= 0 \\ F(x^k) &= \phi(x_E^k) \end{aligned}$$



where  $\phi$  and  $O_E$  are defined in (5.27) and (5.28) respectively.

Since SIPM reduces to projected HBF, parameter choices could be made to optimize the performance of HBF on the optimal manifold. Such computations have been carried out elsewhere [45] for the special case of Problem  $\ell_1$ -LS. The analysis closely follows the original work by Polyak for HBF [36]. For strongly convex quadratic functions, HBF obtains linear convergence with a  $\sqrt{\kappa}$ -times lower rate than gradient descent, where  $\kappa$  is the condition number of the Hessian. For Problem  $\ell_1$ -LS, when  $l_E > 0$  the optimal asymptotic iteration complexity of HBF turns out to be equal to that of I-FBS with our optimal choice given in (5.34). However we can only guarantee local linear convergence for non-decreasing choices of  $\alpha_k$  in the range  $[0, \bar{\alpha}]$  with  $\bar{\alpha}$  less than  $1/3$ . So the locally optimal value of the momentum must be less than  $1/3$  for this to hold. Otherwise convergence is linear with a worse iteration complexity [45].

**6. Numerical Simulations.** We now compare several choices of parameters for I-FBS applied to a random instance of Problem  $\ell_1$ -LS. To compute  $E$  and thus find the locally optimal parameter choice given in (5.31), we use the interior point solver of [46] to find a solution to a target duality gap of  $10^{-8}$ . We then compute  $h^* = \nabla f(x^*)$ , and approximate the set  $E$  by the set of all entries such that  $\rho - |h_i^*|$  is smaller than  $10^{-4}$ . We also use the interior point solver to find an estimate of  $F^*$ . Recall that  $l_E$  denotes the smallest eigenvalue of  $A_E^T A_E$  and note that  $l_E$  was greater than 0 in all experiments we ran. The parameter choices under consideration will be referred to by the following designations.

- **I-FBS-OPT.** I-FBS with our locally optimal parameters derived in Section 5.4, given in (5.31), with  $\lambda = 1/L$ . Recall that this choice optimizes the local convergence rate. Note that this is not a practically implementable algorithm as it depends on the optimal momentum. We include it to verify the theory of Section 5.
- **ISTA-OPT [17].** ISTA with parameters chosen to optimize the local convergence rate. Corresponds to I-FBS with  $\alpha_k = 0$  and step-size  $\lambda_k = 2/(L + l_E)$ . As with I-FBS-OPT this step-size cannot be used in practice as it depends on  $l_E$ . Using  $\lambda_k = 1/L$  is common in practice but gives worse convergence rate.
- **FISTA-BT [16].** FISTA with Beck and Teboulle's parameters given in (3.1).
- **FISTA-AdOPT.** Our proposed method, see Section 5.6. FISTA with Beck and Teboulle's choice until  $F(x^k) > F(x^{k-1})$ , (equivalently  $(y^{k+1} - x^{k+1})^T (x^{k+1} - x^k) > 0$ ), I-FBS-OPT for all iterations thereafter. We estimate the locally optimal momentum by using  $\text{supp}(x^k)$  as a surrogate for  $E$ , since  $\text{supp}(x^k) \subset E$  for  $k$  large enough. We compute the smallest eigenvalue of  $A_{\text{supp}(x^k)}^T A_{\text{supp}(x^k)}$  one time when  $F(x^k) > F(x^{k-1})$  and compute the locally optimal momentum using (5.31). We use this fixed value of the momentum from then on. This algorithm is practically implementable so long as  $|\text{supp}(x^k)|$  is not too large.
- **FISTA-AdRe [34].** Adaptive momentum restart for FISTA. Beck and Teboulle's parameter choice given in (3.1) except set  $t_k$  to 0 whenever  $F(x^k) > F(x^{k-1})$  (equivalently  $(y^{k+1} - x^{k+1})^T (x^{k+1} - x^k) > 0$ ). Our analysis shows this method achieves the optimal asymptotic iteration complexity derived in Corollary 5.3 (See the remarks after Corollary 5.4).

All algorithms are initialized to  $x^1 = x^0 = 0$ .

**Experiment Details.** We create a random instance of Problem  $\ell_1$ -LS, with  $A$  of size  $300 \times 2000$  and with entries drawn i.i.d. from  $\mathcal{N}(0, 0.01)$ . The vector  $b$  is  $Ax_0$  with  $x_0$  being 50-sparse having non-zero entries drawn i.i.d.  $\mathcal{N}(0, 1)$ . The regularization

parameter  $\rho$  is set to 1.

The results are shown in Fig. 6.1. Both FISTA-AdOPT and FISTA-AdRe inherit the  $O(1/k^2)$  convergence rate of FISTA-BT during the transient period. However they also have the optimal asymptotic rate of I-FBS-OPT. FISTA-BT begins to oscillate once the optimal manifold is identified and the iteration complexity is worse than I-FBS-OPT, as predicted in Section 5.5. The performances of FISTA-AdOPT and FISTA-AdRe are similar. FISTA-AdOPT has the advantage that it will not oscillate like FISTA-AdRe, which has to be continually reset once the momentum exceeds the optimal value. However FISTA-AdOPT requires computation of an estimate of the locally optimal momentum which depends on smallest eigenvalue of  $A$  restricted to the support.

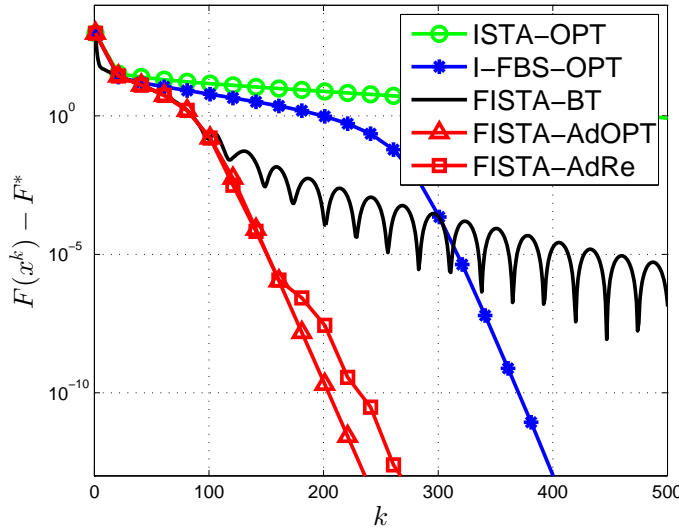


FIG. 6.1. *Experiment results: showing  $F(x^k) - F^*$  versus iteration  $k$  for several algorithms.*

**7. Conclusions.** In this paper, we applied a Lyapunov analysis to a family of inertial forward-backward splitting methods for convex composite minimization. We have proved weak convergence under the following broad conditions with the standard assumptions on the objective function: for the momentum parameter,  $0 \leq \alpha_k \leq 1$  for all  $k$  and  $\limsup \alpha_k < 1$ , and for the step-size, non-decreasing and  $0 < \lambda_k \leq 1/L$ . These conditions are more general than the specific sequences studied in [16], [26] and [27] and less restrictive than the conditions derived in [28]. We considered in detail the behavior of I-FBS applied to sparse optimization problems. With the aid of some results from the Lyapunov analysis we were able to show that I-FBS achieves local linear convergence for these problems, with finite convergence of certain quantities. The local linear convergence results also apply to the FISTA-like methods of [16], [26] and [27].

An interesting direction of future research is to see if this local linear convergence behavior holds for a more general class of problems satisfying certain properties such as partial smoothness and local strong convexity, as considered in [30] for the forward-backward algorithm.

**Acknowledgments.** We would like to thank Olgica Milenkovic, Angelia Nedić, Amin Emad, Antonin Chambolle and Charles Dossal for helpful discussions.

## REFERENCES

- [1] Kwangmoo Koh, Seung-Jean Kim, and Stephen P Boyd, “An interior-point method for large-scale  $\ell_1$ -regularized logistic regression,” *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.
- [2] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani, *The elements of statistical learning*, vol. 2, Springer, 2009.
- [3] Emmanuel J. Candès and Terence Tao, “Decoding by linear programming,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [4] David L Donoho, “Compressed sensing,” *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [5] Leonid I Rudin, Stanley Osher, and Emad Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.
- [6] Antonin Chambolle, “An algorithm for total variation minimization and applications,” *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, 2004.
- [7] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al., “Least angle regression,” *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [8] Cédric Vonesch and Michael Unser, “A fast multilevel algorithm for wavelet-regularized image restoration,” *Image Processing, IEEE Transactions on*, vol. 18, no. 3, pp. 509–523, 2009.
- [9] Stanley Osher, Martin Burger, Donald Goldfarb, Jinjun Xu, and Wotao Yin, “An iterative regularization method for total variation-based image restoration,” *Multiscale Modeling & Simulation*, vol. 4, no. 2, pp. 460–489, 2005.
- [10] Emmanuel J. Candès and Benjamin Recht, “Exact matrix completion via convex optimization,” *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–772, Dec. 2009.
- [11] Pierre-Louis Lions and Bertrand Mercier, “Splitting algorithms for the sum of two nonlinear operators,” *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.
- [12] Gregory B Passty, “Ergodic convergence to a zero of the sum of monotone operators in Hilbert space,” *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2, pp. 383–390, 1979.
- [13] Patrick L Combettes and Jean-Christophe Pesquet, “Proximal splitting methods in signal processing,” in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.
- [14] Boris T Polyak, *Introduction to Optimization*, Optimization Software Inc., 1987.
- [15] Neal Parikh and Stephen Boyd, “Proximal algorithms,” *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.
- [16] Amir Beck and Marc Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009.
- [17] Elaine T. Hale, Wotao Yin, and Yin Zhang, “Fixed-point continuation for  $\ell_1$ -minimization: methodology and convergence,” *SIAM J. on Optimization*, vol. 19, no. 3, pp. 1107–1130, Oct. 2008.
- [18] Kristian Bredies and Dirk A Lorenz, “Linear convergence of iterative soft-thresholding,” *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 813–837, 2008.
- [19] Jonathan Eckstein and W Yao, “Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results,” *RUTCOR Research Reports*, vol. 32, 2012.
- [20] Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr, “An accelerated linearized alternating direction method of multipliers,” *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 644–681, 2015.
- [21] Laurent Condat, “A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms,” *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.
- [22] Ernie Esser, “Applications of Lagrangian-based alternating direction methods and connections to split Bregman,” *CAM report*, vol. 9, pp. 31, 2009.
- [23] Hugo Raguet, Jalal Fadili, and Gabriel Peyré, “A generalized forward-backward splitting,” *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.
- [24] Yurii Nesterov, “A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ ,” in *Soviet Mathematics Doklady*, 1983, vol. 27, pp. 372–376.
- [25] Yurii Nesterov, *Introductory lectures on convex optimization: a basic course*, Springer, 2004.

- [26] Paul Tseng, “On accelerated proximal gradient methods for convex-concave optimization. Submitted to SIAM J. Opt.,” 2008.
- [27] Antonin Chambolle and Charles Dossal, “On the convergence of the iterates of FISTA,” *Preprint hal-01060130*, September, 2014.
- [28] Dirk A. Lorenz and Thomas Pock, “An inertial forward-backward algorithm for monotone inclusions,” *Journal of Mathematical Imaging and Vision*, pp. 1–15, 2014.
- [29] Paul-Emile Maingé, “Convergence theorems for inertial KM-type algorithms,” *Journal of Computational and Applied Mathematics*, vol. 219, no. 1, pp. 223–236, 2008.
- [30] Jingwei Liang, Jalal Fadili, and Gabriel Peyré, “Local linear convergence of forward-backward under partial smoothness,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1970–1978.
- [31] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright, “Fast global convergence rates of gradient methods for high-dimensional statistical recovery,” in *Advances in Neural Information Processing Systems*, 2010, pp. 37–45.
- [32] Laurent Demanet and Xiangxiong Zhang, “Eventual linear convergence of the Douglas-Rachford iteration for basis pursuit,” *arXiv preprint arXiv:1301.0542*, 2013.
- [33] Shaozhe Tao, Daniel Boley, and Shuzhong Zhang, “Local Linear Convergence of ISTA and FISTA on the LASSO Problem,” Tech. Rep., University of Minnesota, 2015.
- [34] Brendan O’Donoghue and Emmanuel Candès, “Adaptive restart for accelerated gradient schemes,” *Foundations of Computational Mathematics*, pp. 1–18, 2012.
- [35] A. Moudafi and M. Oliny, “Convergence of a splitting inertial proximal method for monotone operators,” *Journal of Computational and Applied Mathematics*, vol. 155, no. 2, pp. 447 – 454, 2003.
- [36] Boris Teodorovich Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.
- [37] Boushra Abbas and Hedy Attouch, “Dynamical systems and forward-backward algorithms associated with the sum of a convex subdifferential and a monotone cocoercive operator,” *arXiv preprint arXiv:1403.6312*, 2014.
- [38] Heinz H Bauschke and Patrick L Combettes, “The Baillon-Haddad theorem revisited,” *arXiv preprint arXiv:0906.0807*, 2009.
- [39] Heinz H Bauschke and Patrick L Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer Science & Business Media, 2011.
- [40] Paul-Emile Maingé, “Inertial iterative process for fixed points of certain quasi-nonexpansive mappings,” *Set-Valued Analysis*, vol. 15, no. 1, pp. 67–79, 2007.
- [41] Qihang Lin and Lin Xiao, “An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization,” *Computational Optimization and Applications*, vol. 60, no. 3, pp. 633–674, 2014.
- [42] Felipe Alvarez, “On the minimizing property of a second order dissipative system in Hilbert spaces,” *SIAM Journal on Control and Optimization*, vol. 38, no. 4, pp. 1102–1119, 2000.
- [43] Zdzisław Opial, “Weak convergence of the sequence of successive approximations for nonexpansive mappings,” *Bulletin of the American Mathematical Society*, vol. 73, no. 4, pp. 591–597, 1967.
- [44] Felipe Alvarez and Hedy Attouch, “An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping,” *Set-Valued Analysis*, vol. 9, no. 1-2, pp. 3–11, 2001.
- [45] Patrick R Johnstone and Pierre Moulin, “Convergence of an inertial proximal method for L1-regularized least-squares,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, available at <http://web.engr.illinois.edu/projects2/pdfs/Johnstone.pdf>. IEEE, April 2015.
- [46] S-J Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd, and Dmitry Gorinevsky, “An interior-point method for large-scale  $\ell_1$ -regularized least squares,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 606–617, 2007.